A Teaching Database for Diagnosis of Hematologic Neoplasms Using Immunophenotyping by Flow Cytometry

Andy N. D. Nguyen, MD; Jitakshi De, MD; Jacqueline Nguyen, DO; Anthony Padula, MD; Zhenhong Qu, MD, PhD

• *Context.*—In the diagnosis of lymphomas and leukemias, flow cytometry has been considered an essential addition to morphology and immunohistochemistry. The interpretation of immunophenotyping results by flow cytometry involves pattern recognition of different hematologic neoplasms that may have similar immunologic marker profiles. An important factor that creates difficulty in the interpretation process is the lack of consistency in marker expression for a particular neoplasm. For this reason, a definitive diagnostic pattern is usually not available for each specific neoplasm. Consequently, there is a need for decision support tools to assist pathology trainees in learning flow cytometric diagnosis of leukemia and lymphoma.

***Objective.*—Development of a Web-enabled relational database integrated with decision-making tools for teaching flow cytometric diagnosis of hematologic neoplasms.**

***Design.*—This database has a knowledge base containing patterns of 44 markers for 37 hematologic neoplasms. We have obtained immunophenotyping data published in the scientific literature and incorporated them into a mathematical algorithm that is integrated to the database for differential diagnostic purposes. The algorithm takes into account the incidence of positive and negative expression of each marker for each disorder.**

***Results.*—Validation of this algorithm was performed using 92 clinical cases accumulated from 2 different medical centers. The database also incorporates the latest World Health Organization classification for hematologic neoplasms.**

***Conclusions.*—The algorithm developed in this database shows significant improvement in diagnostic accuracy over our previous database prototype. This Web-based database is proposed to be a useful public resource for teaching pathology trainees flow cytometric diagnosis.**

(*Arch Pathol Lab Med.* 2008;132:829–837)

The use of flow cytometry in immunophenotyping has undoubtedly added an essential dimension to the diagnosis of hematologic neoplasms.[1,2] Hematologic cells express a wide range of cell surface and cytoplasmic antigens.[2–4] The detection of these antigens by flow cytometry allows identification of immunophenotypic profiles associated with lymphomas and leukemias. Many studies have been done to identify the marker patterns of different lymphomas and leukemias. However, a definitive diagnostic pattern is usually not present for each specific neoplasm. Instead, the usual diagnostic approach is to seek a neoplasm that best fits the marker expression profile derived from flow cytometry or immunohistochemistry.[1,2] A well-known problem in the interpretation of immunophenotyping results is the inconsistency in markers expressed in a particular neoplasm.[2–4] A certain marker may be positive (or negative) in most of the cases. However, aberrant expression of malignant cells often introduces exceptions to this typical finding. While these nuances should not present any significant difficulty to an experienced hematopathologist, they typically create problems for pathology trainees. Subsequently, there is a need for decision support tools to teach pathology trainees in diagnosing leukemia and lymphoma using flow cytometry data. We propose that a more accurate and consistent diagnostic pattern could be obtained from a large number of previously diagnosed cases in the scientific literature by taking into account the incidence (relative frequency) of the typical antigen expression.

Since the immunophenotypic pattern of hematologic neoplasms can easily be described in terms of the presence or absence of markers included in a panel, a database is a logical approach to facilitate the representation and interpretation of marker results. We describe research aimed at designing and validating a Web-based database, named CD-MarkerPF, for refining the diagnostic criteria of hematologic neoplasms using results of immunophenotyping by flow cytometry. Our specific aims are:

1. To design a database that assists pathology trainees in the diagnosis of hematologic neoplasms. This database incorporates the latest World Health Organization (WHO) classification for hematologic neoplasms. A mathematical algorithm is integrated into the database to refine the di-

Accepted for publication October 29, 2007.

From the Department of Pathology and Laboratory Medicine, University of Texas Health Science Center at Houston (Drs A. N. D. Nguyen, De, J. Nguyen, and Padula); and the Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester, NY (Dr Qu).

The authors have no relevant financial interest in the products or companies described in this article.

Reprints: Andy N. D. Nguyen, MD, Department of Pathology and Laboratory Medicine, University of Texas–Houston, Medical School, 6431 Fannin St, MSB 2.292, Houston, TX 77030 (e-mail: Nghia.D.Nguyen@uth.tmc.edu).

agnostic criteria for hematologic neoplasms. We use immunophenotyping results published in scientific journal articles. The algorithm takes into account the incidence of positive and negative results of each marker for each disorder. Validation of this algorithm is performed using 92 clinical cases accumulated at two large medical centers.

2. To establish this Web-based database as a public resource for teaching purposes. Pathology trainees can get access to this database on the Internet when needed in their training for the most up-to-date information.

Previous Study: CD-MarkerDX

This database prototype, the predecessor of the current database, represented an innovative application of medical informatics to teaching laboratory diagnosis of leukemia and lymphoma. A total of 33 hematologic neoplasms and 42 immunologic markers were included in database CD-MarkerDX.[5] The diagnostic criteria for different neoplasms were based on the pattern of immunologic marker results. The marker result was designated as positive (or negative) for a neoplasm if more than 50% of the cases were found to be positive (or negative) for that marker as observed in the scientific literature.

A list of differential diagnoses is provided by CD-MarkerDX with each set of input data. The differential diagnoses have an assigned value of matching factor (*MF*). The *MF* value for a neoplasm reflects how well its immunophenotyping pattern matches the marker data in a given case. This factor is defined as[5]:

$$MF = \frac{M}{M + N} \qquad (1)$$

where *MF* indicates the matching factor for a particular neoplasm ($0 \leq MF \leq 1$); *M*, the number of attributes of a neoplasm that match the input data; and *N*, the number of attributes of a neoplasm that do not match the input data.

The value of $(M - N)$ is used as a secondary criterion in ranking differential diagnoses with the same *MF* value. We tested this database using 92 clinical cases from 2 tertiary medical centers. The database ranked the actual diagnosis as one of the top 5 differential diagnoses in 93% of the cases tested.

Note that in this database prototype, a marker is defined as either positive or negative for a certain disorder for simplicity. It was not designed to take into account the incidence of positivity or negativity for each of the markers in each disorder. In the current project, the marker incidence is incorporated in the algorithm and is a major improvement of the designed database.

MATERIALS AND METHODS

The overall goal of this study is to develop a Web-based relational database to assist pathology trainees in learning diagnosis of hematologic neoplasms using immunophenotyping data by flow cytometry. A built-in algorithm in this database is designed to refine the diagnostic criteria of hematologic neoplasms. This Web-based database is implemented as a public resource for training purposes.

Database Design

Relational databases with comprehensive content information and efficient query mechanisms can perform effectively as decision support systems to help users solve complicated problems in laboratory diagnosis. Such systems have 2 major elements: the development environment and the consultation environment (Figure 1).

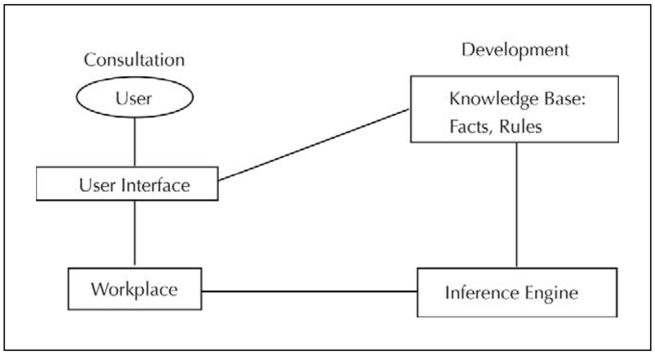


Figure 1. *Architecture of database as a decision support system. A user can get access to the database through user interface to input data and obtain results. The inference engine queries the knowledge base to obtain a set of differential diagnoses.*

The development environment is used by the database builder to construct the database components and to enter information into the knowledge base. The consultation environment is accessed by the user to obtain recommendations. The following components are implemented in our database:

1. Knowledge base. This component contains the knowledge required for formulating and solving problems. It contains facts in the domain area and rules that direct the use of facts to diagnose specific disorders. The diagnostic criteria are the facts, also known as attributes, that are necessary to confirm a certain disorder. Potential sources of knowledge include human experts and scientific literature.
2. Workplace. This component is an area in the computer's working memory for the description of a current problem, as specified by the input data. The workplace also stores intermediate conclusions.
3. Inference engine. This is the central element of the database that provides the methodology for query by using information in the knowledge base and in the workplace to formulate conclusions using mathematical algorithms. The inference engine is designed in the form of a database management system (DBMS), an application that accesses information stored in a database.
4. User interface. This component allows communication between the user and the DBMS. The user uses this interface to input data (positive and negative attributes found in a patient) and to obtain the results. This communication interface is typically in a graphics format for ease of use (graphic user interface).

A total of 37 types of hematologic neoplasms are included in the database (Table 1). The diagnostic criteria for different neoplasms are based on the pattern of immunologic marker results. A total of 44 most commonly used immunologic markers are used to characterize the diagnostic pattern of each neoplasm (Table 2). The marker panel includes only the markers deemed to be most commonly used. As other markers become more extensively used with their added value in diagnosis, they can be incorporated into the database. Also note that the following markers are mostly available as immunoperoxidase stains: cytokeratin, Bcl-1, Bcl-2, and Bcl-6. They are intentionally included in the marker panel to increase the diagnostic accuracy. Immunoperoxidase stains are often used in clinical practice to supplement flow cytometry markers in difficult cases. Cytochemical stains are not included in the marker panel of the DBMS. We used PubMed to search for published data on immunophenotypes of leukemia and lymphoma by flow cytometry. Our attempt for such literature search yielded journal articles with comprehensive data on incidence (frequently expressed as percentage or ratio) of positive and negative marker results for various leukemias and lymphomas. The data on incidence are incorporated in the differential

Table 1. List of Disorders in Database*

Acute myeloblastic leukemia minimally differentiated, M0
Acute myeloblastic leukemia without maturation, M1
Acute myeloblastic leukemia with maturation, M2
Acute promyelocytic leukemia, M3
Acute myelomonocytic leukemia, M4
Acute monocytic leukemia, M5
Acute erythroleukemia, M6
Acute megakaryoblastic leukemia, M7
Biphenotypic acute leukemia, AML + precursor T ALL
Biphenotypic acute leukemia, AML + precursor B ALL
Precursor T lymphoblastic leukemia/lymphoma
Precursor B lymphoblastic leukemia/lymphoma
Thymoma
Follicular lymphoma
Mantle cell lymphoma
Diffuse large B-cell lymphoma
Mediastinal large B-cell lymphoma
Plasma cell neoplasms
Burkitt lymphoma/leukemia
Splenic lymphoma with villous lymphocytes
Primary effusion lymphoma
Chronic lymphocytic leukemia/small lymphocytic lymphoma
B-cell prolymphocytic leukemia
Hairy cell leukemia
Lymphoplasmacytic lymphoma
Marginal zone lymphoma
Plasmablastic lymphoma
Sezary syndrome/mycosis fungoides
Adult T-cell leukemia/lymphoma
NK cell large granular lymphocytosis
T-cell large granular lymphocytic leukemia/lymphoma
T-cell prolymphocytic leukemia
Blastic NK cell lymphoma
Nasal NK T-cell lymphoma/aggressive NK cell leukemia-lymphoma
Enteropathy-type T-cell lymphoma
Subcutaneous panniculitis-like T-cell lymphoma
Peripheral T-cell lymphoma

* AML indicates acute myeloid leukemia, ALL, acute lymphoblastic leukemia; and NK, natural killer.

Table 2. List of Immunologic Markers in Database*

CD1
CD2
CD3
CD4
CD5
CD7
CD8
CD10
CD11b
CD11c
CD13
CD14
CD15
CD16
CD19
CD20
CD21
CD22
CD23
CD24
CD25
CD33
CD34
CD38
CD41
CD42
CD43
CD45
CD56
CD57
CD61
CD71
CD79a
CD103
HLA-DR
sIg
cIg
TdT
FMC7
Glycophorin A
Cytokeratin
Bcl-1
Bcl-2
Bcl-6

* sIg indicates surface immunoglobulin; cIg, cytoplasmic immunoglobulin; and TdT, terminal deoxynucleotidyl transferase.

diagnosis process (see ''Mathematical Algorithm''). The DBMS main menu has 4 modules:

1. Differential diagnosis: to generate a list of differential diagnoses that closely match the marker results in a given case;
2. Display of disorders: typical results of markers for each disorder;
3. Display of markers: relevant information on each immunologic marker;
4. WHO Hematopathology Classification Review: synopsis of hematologic neoplasms.

Software Platform

The following are 2 main components of the software platform used in this project: (1) Microsoft .NET Framework (Microsoft, Seattle, Wash), a computing platform that simplifies application development in the highly distributed environment of the Internet[6]; and (2) C# (pronounced C sharp), a relatively new programming language designed for building a wide range of enterprise applications that run on the .NET Framework.[7]

A major challenge in designing dynamic databases on the Web has been the accommodation of various types of Web browsers that use different client-side technology (Client-side Active X, different Java versions, etc). The .NET Framework facilitates development of browser-independent databases on the World Wide Web.[8,9] All of the processing work is done on the Web server, allowing for the use of a ''thin'' client (a Web browser without any plug-ins or extensions). This database and its associated DBMS are installed on a Microsoft Windows XP server running Microsoft Internet Information Server 6.0. The data reside in a Microsoft SQL Server 2000.

Mathematical Algorithm

A list of differential diagnoses is provided by the DBMS with each set of input data. The differential diagnoses have an assigned value of profile factor (*PF*). The *PF* value for a neoplasm reflects how well its immunophenotyping pattern matches that of a given case. In this project, accumulated data from published literature are used to enhance the sensitivity and specificity of the differential diagnoses. The formula for calculating *PF* is

$$PF = \frac{\sum C_n}{N} \tag{2}$$

where *PF* indicates the profile factor for a particular neoplasm ($0 \le PF \le 1$); C_n, profile coefficient for an input data ($0 \le C_n \le 1$); $n = 1$ to N; and N, the number of (not-NULL) attributes of a neoplasm that have input data.

The profile coefficient is calculated as: $C_n = PosRatio_{(i,j)}$ if input data are positive (+), and $NegRatio_{(i,j)}$ if input data are negative (−).

The ratio of cases that are positive for a certain marker for a disorder, $PosRatio_{(i,j)}$, is defined as:

$$PosRatio_{(i,j)} = \frac{\sum PosCase_{(i,j)}}{\sum Case_{(i,j)}} \tag{3}$$

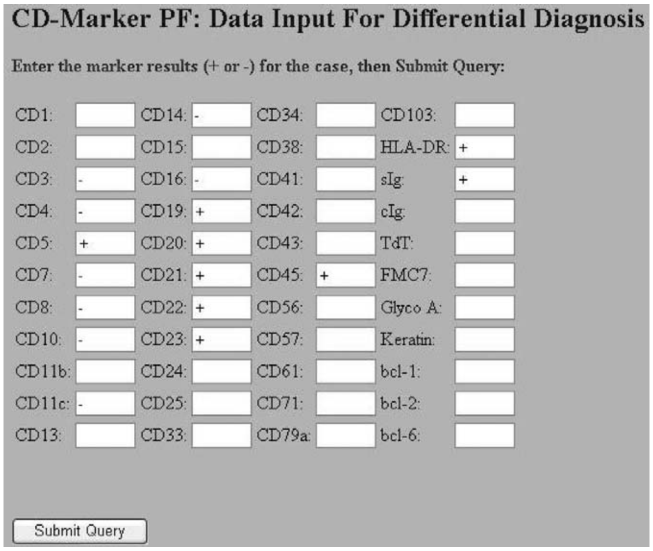
where i indicates the ith disorder; j, the jth marker; $PosCase_{(i,j)}$, number of cases that are positive for the jth marker for the ith disorder; and $Case_{(i,j)}$, the number of cases under study for the jth marker for the ith disorder.

The ratio of cases that are negative for a certain marker for a disorder, $NegRatio_{(i,j)}$, is calculated as:

$$NegRatio_{(i,j)} = 1 - PosRatio_{(i,j)} \tag{4}$$

Note that the calculation of *PF* (Equation 2) is derived from that of *MF* (equation 1) developed previously for our DBMS prototype, CD-MarkerDX. The value of *PF* is expected to be more accurate for diagnostic purposes since it takes into account retrospective data from a large number of previously diagnosed cases.

The importance of certain critical markers in diagnosing a disorder is also considered. If a certain marker is very specific for a disorder, its contributing weight is considered twice as much as other markers in the disorder's profile. In the database table, a



CD-Marker PF: Data Input For Differential Diagnosis

Enter the marker results (+ or -) for the case, then Submit Query:

CD1:		CD14:	-	CD34:		CD103:	
CD2:		CD15:		CD38:		HLA-DR:	+
CD3:	-	CD16:	-	CD41:		sIg:	+
CD4:	-	CD19:	+	CD42:		cIg:	
CD5:	+	CD20:	+	CD43:		TdT:	
CD7:	-	CD21:	+	CD45:	+	FMC7:	
CD8:	-	CD22:	+	CD56:		Glyco A:	
CD10:	-	CD23:	+	CD57:		Keratin:	
CD11b:		CD24:		CD61:		bcl-1:	
CD11c:	-	CD25:		CD71:		bcl-2:	
CD13:		CD33:		CD79a:		bcl-6:	

Figure 2. *Screen shot of data input for a consultation session. Data input is entered by user into text boxes in the browser interface. A plus sign (+) or a minus sign (−) is entered if the marker result is positive or negative, respectively. A text box will be left blank if the result is not available for that particular marker. User will then click on the button "Submit Query" to obtain a list of differential diagnoses (shown in Figure 3).*

negative sign (−) is used to denote this specificity of the marker for a certain disorder.

$PosRatio_{(i,j)}$ would be calculated as $-2 \times PosRatio_{(i,j)}$; $NegRatio_{(i,j)}$ would be calculated as $1 + 2 \times NegRatio_{(i,j)}$

For differential diagnosis of a given case, all of the data that are available on marker results should be entered for the case under consideration. Lack of information in certain data fields does not prevent the DBMS from processing the data. However, the accuracy of the suggested diagnosis would be compromised if results of important markers were left out. When the profile of a given case is entered, it is processed by the DBMS inference engine, and a list of differential diagnoses will be displayed. These diagnoses are listed with their associated *PF* value.

A demonstration of a case with chronic lymphocytic leukemia/small lymphocytic lymphoma (CLL/SLL) is illustrative of how the DBMS inference engine can interpret the immunophenotyping results and how its search mechanism works (Figure 2). The marker data available for the patient sample are

1. Positive for CD5, CD19, CD20, CD22, CD23, CD45, HLA-DR, and surface immunoglobulin (sIg).
2. Negative for CD3, CD4, CD7, CD8, CD10, CD11c, CD14, and CD16.

The DBMS inference engine attempts to match this set of data with the diagnostic attributes of 37 hematologic neoplasms in the knowledge base. The total number of attributes of CLL/SLL that have input data is 16 (8 positive results and 8 negative results). This number is represented by the variable *N* in equation 2 ($N = 16$). Note that the following attributes in the knowledge base for CLL/SLL did not have a corresponding data input: CD1, CD2, CD11b, CD13, CD15, CD21, CD24, CD25, CD33, CD34, CD38, CD41, CD42, CD43, CD56, CD57, CD61, CD71, CD79a, CD103, cytoplasmic immunoglobulin (cIg), terminal deoxynucleotidyl transferase (TdT), FMC7, glycophoryn A, keratin, Bcl-1, Bcl-2, and Bcl-6. These attributes do not have any impact on the ranking of CLL/SLL since they are not included as part of the calculation for its *PF.* The intentional exclusion of attributes without corresponding input data in calculating *PF* serves an important purpose of maintaining a flexible design for the knowledge base as well as for the data input panel. Since different flow cytometry laboratories may use different markers in immunophenotyping, and various studies on marker pattern of neoplasms have used different marker panels, an absolute requirement of certain markers in the interpretation process would be too stringent to yield any reasonable matches.[10]

The following calculations are performed to determine the value of *PF* for CLL/SLL:

$$\text{Sum of } PosRatio_{(i,j)}$$

$$= [2(0.9) + 0.95 + 0.95 + 0.6 + 2(0.9) + 1.0 + 0.9 + 1.0] = 9$$

$$\text{Sum of } NegRatio_{(i,j)}$$

$$= [1 + 1 + 1 + 1 + 1 + 0.33 + 1 + 1] = 7.33$$

$$PF = \frac{\sum_{n} C_n}{N} = \frac{9 + 7.33}{16} = 1.02$$

Note that the value of C_n is equal to $PosRatio_{(i,j)}$ for the first 8 markers (with positive input) and is equal to $NegRatio_{(i,j)}$ for the last 8 markers (with negative input).

After the DBMS inference engine calculates the *PF* value for all of the remaining 36 hematologic neoplasms in the knowledge base and ranks them accordingly, it lists the following leading diagnoses (Figure 3):

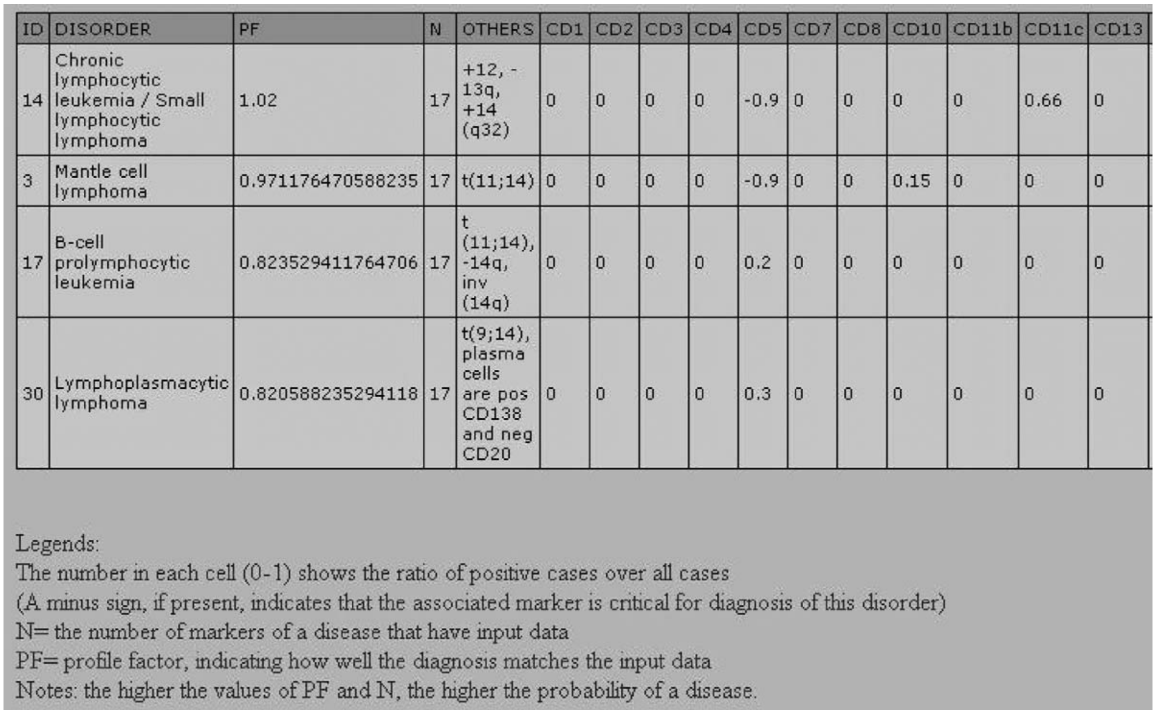
1. CLL/SLL: $PF = 1.021$
2. Mantle cell lymphoma: $PF = 0.969$
3. B-cell prolymphocytic leukemia: $PF = 0.875$
4. Lymphoplasmacytic lymphoma: $PF = 0.871$

Only neoplasms with $PF > 0.800$ are displayed in the list of leading diagnoses. This threshold seems to work well in most cases. A lower or higher threshold can be set in the database. However, this may yield a list of leading diagnoses that is either too short or too long, respectively. The search mechanism of going from neoplasms in the database to the input data for the best matches represents a strategy known as backward chaining.[11–13] This demonstration shows the open-ended format of the data input. The data panel consists of many immunologic markers, some of which may not be part of routine testing in a particular laboratory. Consequently, the actual data input for a case are unlikely to account for all the markers in the data panel. However, the availability of essential data would influence the accuracy of ranking by the database.

The critical role of the interpreting trainee cannot be overemphasized. The DBMS is only useful in suggesting a list of differential diagnoses. The trainee must establish the final diagnosis by correlating the histologic findings of the case with the immunophenotyping results. The immunologic marker patterns of neoplasms in the list of differential diagnoses are displayed side by side for comparison. For a quick review of disorders in the list of differential diagnoses, the user can view the 4 most important data in the 4 leftmost columns: the Disorder, *N* (the number of markers of a disease that have input data), *PF* (profile factor, indicating how well the diagnosis matches the input data), and Others (cytogenetics, cytochemical stains, etc). The listing of markers can be viewed by scrolling the screen to view a complete marker panel. Users also have the option of viewing the full panel of each disorder in tabular form using the "Display of Disorders" option in the main menu (Figure 4). The user also has the option to retrieve marker information during a consultation session to obtain more information on the properties of each marker (Figure 5). Additional information on each neoplasm can also be reviewed from the "WHO Hematopathology Classification Review" feature of this database. The synopsis of each neoplasm offers the user further essential information, such as morphology and clinical features, before finalizing the diagnosis (Figure 6).

Validation Method

We used 92 cases with immunophenotyping data representing various hematologic neoplasms to validate the DBMS. These are patient cases from 2 tertiary medical centers. The cases are highly

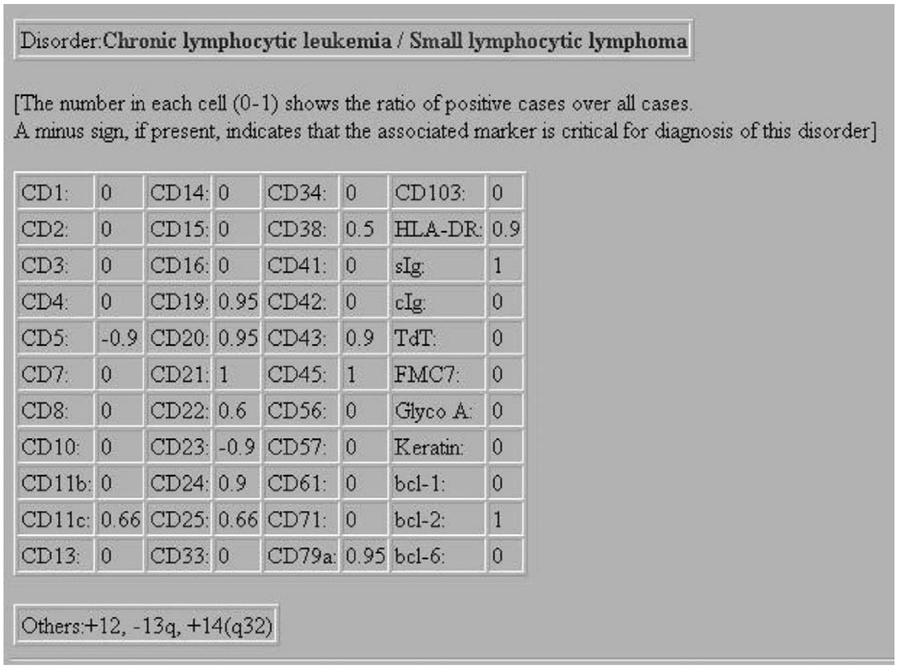


ID	DISORDER	PF	N	OTHERS	CD1	CD2	CD3	CD4	CD5	CD7	CD8	CD10	CD11b	CD11c	CD13
14	Chronic lymphocytic leukemia / Small lymphocytic lymphoma	1.02	17	+12, -13q, +14 (q32)	0	0	0	0	-0.9	0	0	0	0	0.66	0
3	Mantle cell lymphoma	0.971176470588235	17	t(11;14)	0	0	0	0	-0.9	0	0	0.15	0	0	0
17	B-cell prolymphocytic leukemia	0.823529411764706	17	t (11;14), -14q, inv (14q)	0	0	0	0	0.2	0	0	0	0	0	0
30	Lymphoplasmacytic lymphoma	0.820588235294118	17	t(9;14), plasma cells are pos CD138 and neg CD20	0	0	0	0	0.3	0	0	0	0	0	0

Legends:
The number in each cell (0-1) shows the ratio of positive cases over all cases
(A minus sign, if present, indicates that the associated marker is critical for diagnosis of this disorder)
N= the number of markers of a disease that have input data
PF= profile factor, indicating how well the diagnosis matches the input data
Notes: the higher the values of PF and N, the higher the probability of a disease.

Figure 3. *Screen shot of differential diagnoses generated by the database. The input data are compared against all of the 37 neoplasms in the database, and a list of differential diagnoses is displayed. These diagnoses are ranked in decreasing probability. Legends below the list offer explanation of the parameters shown next to each diagnosis. A user can retrieve marker profiles for each neoplasm before making a final diagnosis (shown in Figure 4).*

diversified with regard to sex, age, and ethnic groups. The hematologic neoplastic cases were collected consecutively from January 2004 to December 2005 to ensure no bias in patient selection. Data collection was approved for this project by our institutional review board. Data for these cases were retrospectively retrieved, and immunophenotyping data were entered into the database. The final diagnosis of each case was previously established by histologic findings and correlation with flow cytometry results. The final diagnosis was documented in surgical pathology reports, including bone marrow reports. Data entry of each case in the DBMS included only marker results that were available in the flow cytometry laboratories at the time of initial presentation. A marker was considered positive if expressed by at least 20% of the analyzed cells. Isotopic controls were used to determine autofluorescence and background. Only definitive marker results (positive or negative) in each case were used in validation. Equivocal results were not used due to their lack of contribution to the validation results. The specimens in our cases included bone mar-



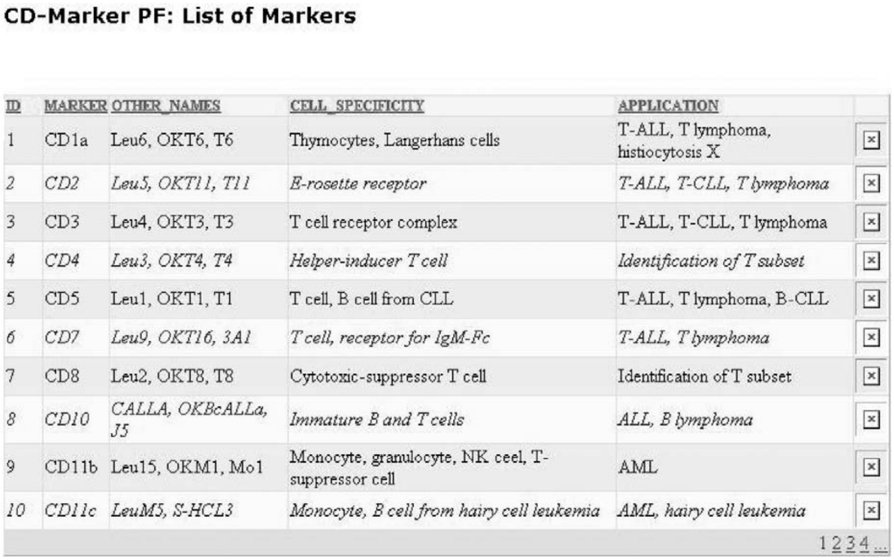
Disorder:**Chronic lymphocytic leukemia / Small lymphocytic lymphoma**

[The number in each cell (0-1) shows the ratio of positive cases over all cases.
A minus sign, if present, indicates that the associated marker is critical for diagnosis of this disorder]

CD1:	0	CD14:	0	CD34:	0	CD103:	0
CD2:	0	CD15:	0	CD38:	0.5	HLA-DR:	0.9
CD3:	0	CD16:	0	CD41:	0	sIg:	1
CD4:	0	CD19:	0.95	CD42:	0	cIg:	0
CD5:	-0.9	CD20:	0.95	CD43:	0.9	TdT:	0
CD7:	0	CD21:	1	CD45:	1	FMC7:	0
CD8:	0	CD22:	0.6	CD56:	0	Glyco A:	0
CD10:	0	CD23:	-0.9	CD57:	0	Keratin:	0
CD11b:	0	CD24:	0.9	CD61:	0	bcl-1:	0
CD11c:	0.66	CD25:	0.66	CD71:	0	bcl-2:	1
CD13:	0	CD33:	0	CD79a:	0.95	bcl-6:	0

Others:+12, -13q, +14(q32)

Figure 4. *Screen shot of marker profile for a disorder in the database. A user has the option to retrieve a marker profile during a consultation session to examine the complete diagnostic criteria for each neoplasm.*



CD-Marker PF: List of Markers

ID	MARKER	OTHER_NAMES	CELL_SPECIFICITY	APPLICATION	
1	CD1a	Leu6, OKT6, T6	Thymocytes, Langerhans cells	T-ALL, T lymphoma, histiocytosis X	☒
2	CD2	Leu5, OKT11, T11	E-rosette receptor	T-ALL, T-CLL, T lymphoma	☒
3	CD3	Leu4, OKT3, T3	T cell receptor complex	T-ALL, T-CLL, T lymphoma	☒
4	CD4	Leu3, OKT4, T4	Helper-inducer T cell	Identification of T subset	☒
5	CD5	Leu1, OKT1, T1	T cell, B cell from CLL	T-ALL, T lymphoma, B-CLL	☒
6	CD7	Leu9, OKT16, 3A1	T cell, receptor for IgM-Fc	T-ALL, T lymphoma	☒
7	CD8	Leu2, OKT8, T8	Cytotoxic-suppressor T cell	Identification of T subset	☒
8	CD10	CALLA, OKBcALLa, J5	Immature B and T cells	ALL, B lymphoma	☒
9	CD11b	Leu15, OKM1, Mo1	Monocyte, granulocyte, NK ceel, T-suppressor cell	AML	☒
10	CD11c	LeuM5, S-HCL3	Monocyte, B cell from hairy cell leukemia	AML, hairy cell leukemia	☒

1 2 3 4 ...

Figure 5. *Screen shot showing part of marker description in the database. A user has the option to retrieve marker information during a consultation session to obtain more information on the properties of each marker.*

row, lymph node, spleen, body fluid, and extranodal hematologic tumors. Immunophenotyping by flow cytometry was performed on FACScan instruments (Becton Dickinson, Mountain View, Calif) as previously described.[1]

The DBMS was validated by the concordance (inclusion and rank) of the actual diagnosis with the differential diagnosis.

Web Site URL

This DBMS can be accessed at the following Web sites (both sites were last accessed on October 25, 2007): http://HemepathReview.com or http://dpalm.med.uth.tmc.edu/faculty/bios/nguyen/Decision.html.

RESULTS

Table 3 shows the results for all of the 92 cases used in this validation process with the accompanying information: ranking of the final diagnosis by the DBMS, the number of cases in each ranking category, and the percentage and accumulated percentage of cases in each ranking category. The validation results show a success rate of 89%. This success rate means that in 89% of the cases, the final diagnosis is included in the list of the top 3 differential diagnoses generated by the DBMS. This represents an improvement over the 80% level by the previous DBMS prototype. The top differential diagnosis shows the actual diagnosis in 60% of the cases by the database, a substantial increase over the 42% level achieved by the previous DBMS prototype.

Note that flow cytometry by itself is insufficient to achieve the final diagnosis. Other data (including morphology, immunohistochemical stains, cytogenetics and, less often, molecular diagnostics) would be needed for a final and correct diagnosis. Expectedly, the top differential diagnosis shown by the system is the actual diagnosis in only 60% of the cases. More importantly, the actual diagnosis is shown in the top 3 differential diagnoses in 89% of the cases, and in the top 5 in 95% of the cases. This pattern of ranking is comparable with the routine process of ruling out disorders in the short list using additional information before getting a final diagnosis.

In 4% of the cases, the final diagnosis was ranked below the top five differential diagnoses for the following reasons:

1. Unusual immunophenotype: a case of CD5-positive, diffuse large B-cell lymphoma (known to be seen in less than 10% of cases).

2. Three cases of T-cell lymphoma: this deficiency is found to be due to an intrinsic limitation of the DBMS in handling certain cases of T-cell malignancies. Aberrant loss of T-cell antigens is a characteristic finding in T-cell malignancies.[1–4] However, a suitable inference mechanism has not been successfully developed to detect such manifestation. The difficulty in designing an algorithm for such detection lies in the random distribution of T-cell markers, making the values of diagnostic attributes impossible to be programmed into the DBMS. Despite this shortcoming, a considerable number of T-cell cases are successfully ranked in the top 5 differential diagnoses by the DBMS (11/14 cases, or 79% of the T-cell cases).

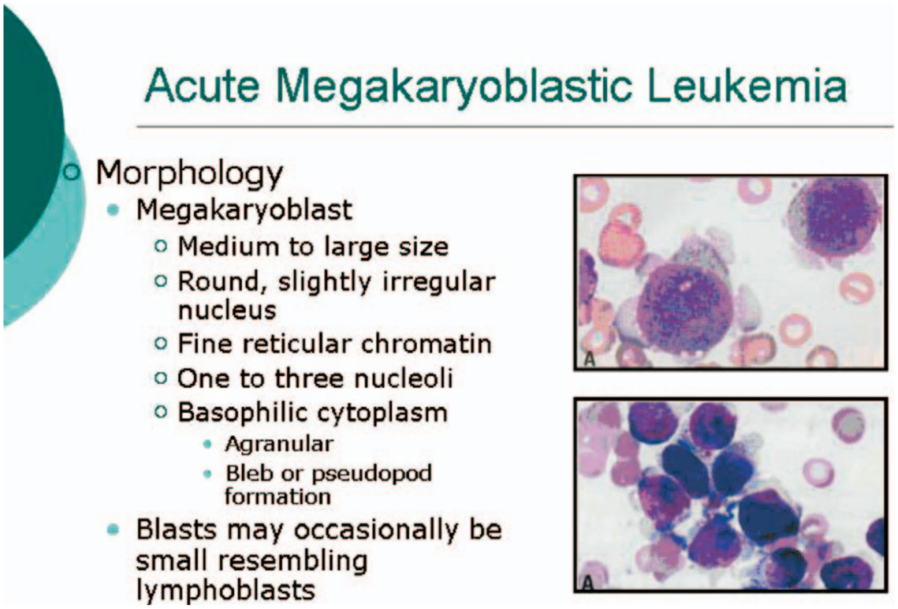
3. Incorrect final diagnosis: a case of lymphoplasmacytic lymphoma was not ranked by the DBMS in the top 5 differential diagnoses. Review of microscopic slides and flow cytometry data for this case indicated that the diagnosis should be revised as CLL/SLL. In fact, CLL/SLL was ranked second by the DBMS.

COMMENT

The cluster designation (CD) of human leukocyte differentiation antigens was formulated by the First and Fourth International Workshops to specify appropriate cell lineage according to the pattern of antigen expression.[3] A wide range of monoclonal antibodies is currently available to recognize various hematologic cells based on their surface and cytoplasmic antigens.[1–4] Leukemic and lympho-

Table 3. Summary of the Validation Results, Rank of the Actual Diagnosis by CD-MarkerPF

Ranking by CD-MarkerPF	No. of Cases	Percentage	Accumulated Percentage
First differential diagnosis	55	60	60
Second differential diagnosis	14	15	75
Third differential diagnosis	13	14	89
Fourth differential diagnosis	4	4	93
Fifth differential diagnosis	2	2	95
Lower ranking	4	4	98
Total	**92**	**100**	. . .



Acute Megakaryoblastic Leukemia

- Morphology
  - Megakaryoblast
    - Medium to large size
    - Round, slightly irregular nucleus
    - Fine reticular chromatin
    - One to three nucleoli
    - Basophilic cytoplasm
      - Agranular
      - Bleb or pseudopod formation
  - Blasts may occasionally be small resembling lymphoblasts

Figure 6. *Screen shot showing part of the synopsis for each neoplasm. A user has the option to retrieve further information during a consultation session to obtain more information on the property of each neoplasm.*

ma cells cannot usually be detected with a single immunologic marker. Instead, the use of an antibody panel consisting of multiple antibodies is required to support the provisional diagnosis based on histologic findings.[1,2] Since many hematologic neoplasms demonstrate similar immunophenotypic patterns, their diagnosis often presents a challenge to pathology residents in training. As the number of immunologic markers used in flow cytometry increases, a systematic approach in interpretation of marker results is also essential for consistent classification of neoplasms.[10]

Clinical decision support systems in the form of computer programs have been hailed for their potential to reduce medical errors and to increase health care quality. At the same time, evidence-based medicine has been widely promoted as a means to improve clinical outcomes. Evidence-based medicine refers to the practice of medicine based on the best available scientific and clinical evidence. The use of clinical decision support systems for teaching evidence-based medicine promises to substantially improve health care quality and efficiency.[14–16] Immunophenotyping of leukemias and lymphomas, with its associated difficulty in interpretation, is one of the ideal areas for teaching evidence-based medicine using decision support systems. To obtain adequate data for this purpose, a comprehensive review of scientific publications is essential to accumulate the incidence of positive and negative result for each marker for each disorder. Review of the literature showed that no such comprehensive study has been attempted. Evidence-based medicine is an essential component of our study to refine the mathematical algorithm used in diagnosis of hematologic neoplasms using flow cytometry.

A number of computer programs have been developed to facilitate interpretation of immunophenotyping for hematologic neoplasms by flow cytometry.[17–19] Different approaches have been attempted, including rule-based systems, cluster analysis, semantic networks, and various mathematical algorithms. They had various degrees of limitation in scope and in accuracy.[5] To use any of these programs, users also have to install them on their computer or computer network. This requirement limits the number of users of these stand-alone programs. Furthermore, no effort has been seen in refining the diagnostic criteria using accumulated data from scientific literature consisting of large numbers of previously diagnosed cases. All of these programs were also based on diagnostic classifications that have become outdated.

The World Wide Web offers a simple solution to the limitation of stand-alone programs by providing easy access to online materials. Existing Internet networks across multiple platforms can be used as the medium for software implementation. Users located in any part of the world with an Internet connection can use browsers to get access to online materials that reside in centralized Web servers. A number of Web sites have been dedicated to different topics in immunophenotyping using flow cytometry.[20–24] Those are usually developed by academic institutions for teaching[20–22] or by commercial vendors for advertising their products.[23,24] While a great deal of valuable information can be retrieved from these sites on many topics, their contents are not focused on teaching programs for interpreting flow cytometry results.

In this project, we successfully designed and validated a decision support tool for teaching pathology trainees the diagnosis of leukemia and lymphoma using flow cytometry data. In doing so, we used imunophenotype data published in scientific literature and incorporated the incidence of positive and negative marker expression for each disorder. A mathematical algorithm was integrated into the DBMS to refine the diagnostic criteria for hematologic neoplasms. Validation of this algorithm was performed using 92 clinical cases accumulated from two different medical centers. The actual diagnosis was ranked as one of the top three differential diagnoses in 89% of the cases, and as the top differential diagnosis in 60% of the cases, a substantial improvement over the 80% and 42% levels with our previous DBMS prototype, respectively.

This database has a knowledge base containing the pattern of 44 markers in 37 hematologic neoplasms based on the latest WHO classification for hematologic neoplasms. This Web-based database is established as a public resource for teaching purpose. All pathology trainees can get access to this DBMS on the Internet when needed in their daily training for the most up-to-date information. Access to the DBMS is readily obtained with any Web browser. No limitation on browser choice is imposed by Active Server Page.NET, which is the basic building block of the DBMS interface. The Web server (Internet Information Server) and the database server (SQL Database Server) also allow for hundreds of users accessing the DBMS at the same time on the Internet. To comply with regulations on patient confidentiality, no patient identification information is stored in the database.

Despite the utility of CD-MarkerPF, there are certain constraints inherent in its use.

1. This DBMS is strictly designed for pathology residents in training. It simply provides the differential diagnosis for a given set of marker expressions without regard to morphology, cytogenetics, clinical presentation, or response to therapy. Therefore, the differential diagnosis is designed to be broad. Just like trained hematopathologists, the trainees must still correlate the immunophenotypic findings with the morphologic characteristics of the neoplasm, the most important basis of diagnosis and classification.
2. The user must have a functional knowledge of hematologic disorders to be able to use CD-MarkerPF effectively, because this DBMS only serves as a search tool to aid the user in making a diagnosis. The technical skills to perform the laboratory procedures and the experience needed to accurately gate the cellular populations are critical in the diagnostic process. CD-MarkerPF can generate a list of differential diagnoses in most cases if adequate data are input. The interpreting trainee can then quickly compare the patient's laboratory data to the marker patterns available from the CD-MarkerPF display module and make the appropriate diagnosis. It cannot be overemphasized that human judgment is the most important element in finalizing the diagnosis.
3. The current version of CD-MarkerPF is deficient in handling some cases of T-cell malignancy due to the difficulty in designing an algorithm for detection of the random loss of T-cell antigens, as discussed earlier.
4. Not all of the commercially available markers were used in our laboratory. Subsequently, our validation results do not represent a maximal accuracy level that would have been achieved if all of the available markers had been used.
5. CD-MarkerPF would not be useful in the diagnosis of neoplasms that traditionally have not been shown to benefit from flow cytometric immunophenotyping, such as classical Hodgkin lymphoma and T-cell lymphomas without loss of T-cell–associated antigens.
6. It is difficult to compare flow cytometry data produced by different laboratories due to the use of different monoclonal antibodies and negative controls, fluorochromes, instrumentations, and specimen processing. This limitation may have accounted for some of the discrepancies in the validation study. Theoretically, it is possible to improve the accuracy of the system by setting up multiple databases to take into account all different combinations of methods in different laboratories (instrumentation, calibration, reagents, and scoring method, etc). However, the resulting system would be extremely complicated and difficult to use by pathology trainees. We resorted to a simplified approach using the available marker results from different studies at face values, with some sacrifice in accuracy.
7. The intensity and uniformity of staining of a certain marker are often useful in interpreting flow cytometry data. One typical example is the dim expression of CD20 in CLL/SLL. Again, for simplification of the user interface we decided to omit these features in the inference process, with some sacrifice in accuracy. Trainees are encouraged to pay close attention to shifts in antigen intensity in neoplastic cells compared with their normal counterparts (eg, CD20, and surface light-chain intensity in CLL cells vs normal peripheral blood B cells).
8. The WHO classification does not specify whether its ''positive'' and ''negative'' descriptions are derived from immunohistochemistry or flow cytometry data. Some of the negative designations in WHO classification are based primarily on immunohistochemistry and may not be applicable to flow cytometry. The risks in using only flow cytometry for immunophenotyping purpose cannot be overemphasized, especially in difficult cases.
9. It has been well known that flow cytometry alone cannot differentiate between M0, M1, and M2. In fact, the ranking of each subtype would be the same for a given case of M0, M1, or M2. Manual differential and cytochemical stains would also be needed for final diagnosis. Instead of lumping these three subtypes of AML together, we kept them separate for a more uniform presentation (M0 to M7).
10. To differentiate between acute myeloid leukemia versus the myelodysplastic syndromes with increased blasts or chronic myeloproliferative diseases in transformation, flow cytometry data need to be interpreted together with manual differential and clinical information.

Undoubtedly, validation of the usefulness of an educational tool such as this DBMS is an important part of the research effort. Currently, the number of hits to our site is moderate, mostly from our trainees and outside trainees who, incidentally, went to our site. We plan to send our Web site address to other pathology-related Web sites to establish links on their sites. Once the number of visitors has increased, a survey will be done for visitors to evaluate our program. We are currently conducting a study at our institutions to correlate the effectiveness of using this online program compared with other learning modalities. We are validating the system as a training tool by assigning real cases to our pathology residents and monitoring their progress by pretests and posttests. It will take a number of years before we have accumulated adequate data to be statistically significant. The results of this study and the online survey will be submitted for publication once they have been completely compiled and analyzed.

This DBMS should not be viewed as a sole tool for interpreting flow cytometric data without knowing the limitations listed above. Our Web site for this DBMS clearly emphasizes this caution in the disclaimer. Under this constraint, CD-MarkerPF is designed to provide a convenient, interactive tool to teach pathology trainees in diagnosing hematologic neoplasms using flow cytometric data. Our program can be considered a prototype for future decision support programs with much more enhanced features in both algorithm and interface. Software development has progressed at a rapid pace and may someday provide useful tools for even the more experienced practitioners.

Our institution supports different platforms, including

.NET, Java, and LINUX. We decided to use .NET and C# mainly because of the available expertise of the corresponding author. Open sources, such as middlewares by JBoss, a division of Red Hat (Raleigh, NC), would be better tools for collaborative work between groups. Even though our .NET platform is not an open source, the code in our program (in C#) can be exported to any open source using Java or C++ with reasonable amount of effort.

References

1. Keren DF, McCoy JJP, Carey JL. *Flow Cytometry in Clinical Diagnosis.* Chicago, Ill: American Society of Clinical Pathologists Press; 2001.

2. Sun T. *Color Atlas and Text of Flow Cytometric Analysis of Hematologic Neoplasms.* New York, NY: Igaku-Shoin; 1993.

3. Kjeldsberg C, ed. *Practical Diagnosis of Hematologic Disorders.* Chicago, Ill: American Society of Clinical Pathologists Press; 2000.

4. Jaffe ES, Harris NL, Stein H, Vardiman JW, eds. *Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues.* Lyon, France: IARC Press; 2001. *World Health Organization Classification of Tumours.*

5. Nguyen A, Uthman MO, Johnson KA. A Web-based database for diagnosis of haematologic neoplasms using immunophenotyping by flow cytometry. *Med Inf Internet Med.* 2001;26:309–323.

6. Esposito D. *Programming Microsoft ASP.NET.* Redmond, Wash: Microsoft Press; 2003.

7. Hejlsberg A, Wiltamuth S, Golde P. *The C# Programming Language.* Redmond, Wash: Microsoft Press; 2004.

8. Willie C. *Unlocking Active Server Pages.* Indianapolis, Ind: New Riders Publishing; 1997.

9. Tanler R. *The Intranet Data Warehouse.* New York, NY: John Wiley & Sons; 1997.

10. Check W. Diagnosing leukemia, lymphoma: when laboratories go with flow analysis. *CAP Today.* 1998:1–37.

11. Chignell M, Parsaye K. *Expert Systems for Experts.* New York, NY: John Wiley & Sons; 1988.

12. Buchanan BG, Shortliffe EH, eds. *Rule-Based Expert Systems.* Reading, Mass: Addison-Wesley; 1984.

13. Turban E. *Expert Systems and Applied Artificial Intelligence.* New York, NY: Macmillan Publishing; 1992.

14. McDonald JM, Brossett S, Moser SA. Pathology information systems: data mining leads to knowledge discovery. *Arch Pathol Lab Med.* 1998;122:409–411.

15. Sim I, Gorman P, Greenes RA, et al. Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc.* 2001;8:527–534.

16. Nguyen A. The application of data mining to flow cytometry. In: Robinson JP, Darzynkiewicz Z, Dean P, eds. *Current Protocols in Cytometry.* New York, NY: John Wiley & Sons; 2002:10.13.1–10.13.6.

17. Alvey PL, Preston NJ, Greaves MF. High performance for expert systems: a system for leukemia diagnosis. *Med Inform.* 1987;12:97–114.

18. Petrovecki M, Marusic M, Dezelic G. An algorithm for leukaemia immunophenotype pattern recognition. *Med Inform.* 1993;18:11–21.

19. Verwer B, Terstappen L. Automatic lineage assignment of acute leukemia by flow cytometry. *Cytometry.* 1993;14:862–875.

20. The Leukemia Information Center. Leukemia library. Available at: http://www.meds.com/leukemia/flow/flow0.html. Accessed September 12, 2007.

21. Golightly DM. Clinical flow laboratory. Available at: http://www.path.sunysb.edu/labsvs/CLINicalFLOW.htm. Accessed September 19, 2007.

22. Robinson DJP. Purdue University cytometry laboratories. Available at: http://www.cyto.purdue.edu/index.htm/. Accessed September 20, 2007.

23. Exalpha Biologicals: Flow Cytometry. Available at: http://www.exalpha.com/FlowCyto.html. Accessed September 13, 2007.

24. Becton Dickinson Biosciences. Flow cytometry products. Available at: http://www.bdbiosciences.com/features/products/. Accessed September 18, 2007.