OPEN

Citation: Blood Cancer Journal (2013) **3**, e129; doi:10.1038/bcj.2013.27
© 2013 Macmillan Publishers Limited All rights reserved 2044-5385/13

www.nature.com/bcj

ORIGINAL ARTICLE

Defining consensus leukemia-associated immunophenotypes for detection of minimal residual disease in acute myeloid leukemia in a multicenter setting

N Feller[1], VHJ van der Velden[2], RA Brooimans[3], N Boeckx[4,5], F Preijers[6], A Kelder[1], I de Greef[7], G Westra[1], JG te Marvelde[2], P Aerts[3], H Wind[2], M Leenders[6], JW Gratama[3] and GJ Schuurhuis[1]

Flow-cytometric detection of minimal residual disease (MRD) has proven in several single-institute studies to have an independent prognostic impact. We studied whether this relatively complex approach could be performed in a multicenter clinical setting. Five centers developed common protocols to accurately define leukemia-associated (immuno)phenotypes (LAPs) at diagnosis required to establish MRD during/after treatment. List mode data files were exchanged, and LAPs were designed by each center. One center, with extensive MRD experience, served as the reference center and coordinator. In quarterly meetings, consensus LAPs were defined, with the performance of centers compared with these. In a learning (29 patients) and a test phase (35 patients), a mean of 2.2 aberrancies/patient was detected, and only 1/63 patients (1.6%) had no consensus LAP(s). For the four centers without (extensive) MRD experience, clear improvement could be shown: in the learning phase, 39–63% of all consensus LAPs were missed, resulting in a median 30% of patients (range 21–33%) for whom no consensus LAP was reported; in the test phase, 27–40% missed consensus LAPs, resulting in a median 16% (range 7–18%) of 'missed' patients. The quality of LAPs was extensively described. Immunophenotypic MRD assessment in its current setting needs extensive experience and should be limited to experienced centers.

Blood Cancer Journal (2013) **3,** e129; doi:10.1038/bcj.2013.27; published online 2 August 2013

Keywords: minimal residual disease; flow cytometry; multicenter

INTRODUCTION

The interest in minimal residual disease (MRD) assessment as a prognostic tool in acute myeloid leukemia (AML) is growing. Molecular markers, although allowing a greater sensitivity, enable MRD monitoring only in a minority of cases.[1] Immunophenotypic MRD detection using flow cytometry turned out to offer an attractive alternative. Using immunophenotyping, several early studies have shown the prospective value of the frequency of MRD cells for survival of AML patients both in adults[2–6] and in children.[7–9] Meanwhile, many other studies have confirmed these data.[1,10] Immunophenotypic detection of MRD using flow cytometry can be performed by defining aberrant expression of markers or marker combinations on the leukemic cells at diagnosis. These so-called leukemia-associated (immuno)phenotypes (LAPs) are absent, or are present at very low frequencies, in normal blood or bone marrow (BM) cells. This approach can be successfully applied in 80–100% of AML patients.[2–9,11]

Apart from serving as a prognostic marker, MRD may ultimately turn out to offer a tool to guide patient-tailored therapy, an approach already adopted once for childhood AML.[12] Even when categorized as low risk by cytogenetics/molecular markers and/or other conventional prognostic parameters, AML patients may be identified to be at high risk based on MRD cell frequency after induction therapy and stratified to treatment intensification such as allogeneic transplantation. Inversely, intensive treatment in AML patients defined as 'poor risk' by conventional factors, but as 'low risk' based on MRD assessment, might ultimately be avoided. Furthermore, at follow-up, increments in MRD cell frequency by sequential MRD assessments may identify forthcoming relapses, in turn urging therapy to be administered before overt relapse occurs.

For these reasons, it is important to establish the prognostic value of MRD cell frequency assessment in clinical studies in prospective, multicenter settings. A first requirement to enable adequate performance of such studies is the standardization of all procedures necessary for definition of adequate LAPs in newly diagnosed AML, as well as those necessary for identification and quantification of MRD in BM follow-up material. Standardizing LAP technology for MRD detection is complicated because aberrant immunophenotypes may not only differ from patient to patient but also within a patient, often resulting in many different LAPs (illustrated in Supplementary Figure 1). There are large differences in the suitability of LAPs. This suitability may be defined by at least four factors. First, the percentage of blasts that bear a particular aberrancy at diagnosis determines the *sensitivity* of subsequent detection and quantification of AML cells. Second, the extent to which equivalent cells in control normal BM bear the aberrancy determines LAP *specificity*: the lower the frequencies of such 'background' expression, the higher the LAP specificity.[5,13,14] Third

[1]Department of Hematology, VU University Medical Center, Amsterdam, The Netherlands; [2]Department of Immunology, Erasmus MC, Erasmus University Medical Center, Rotterdam, The Netherlands; [3]Department of Medical Oncology, Laboratory of Clinical Tumor Immunology, Erasmus MC Daniel den Hoed Cancer Center, Rotterdam, The Netherlands; [4]Department of Laboratory Medicine, UZ Leuven, Leuven, Belgium; [5]Department of Oncology, KUL, Leuven, Belgium; [6]Department of Laboratory Medicine, Laboratory for Hematology, Radboud University Nijmegen Medical Center (RUNMC), Nijmegen, The Netherlands and [7]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. Correspondence: Dr GJ Schuurhuis, Department of Hematology, VU University Medical Center, CCA Building, Room 4.24, De Boelelaan 1117, Amsterdam 1081 HV, The Netherlands. E-mail: gj.schuurhuis@vumc.nl

Received 18 June 2013; accepted 25 June 2013

is the *stability* of LAPs at follow-up—that is, the possibility that expression of markers, which constitute the LAP, may be higher or lower at follow-up as compared with diagnosis,[5,14–16] thereby resulting in overestimation or false negativity, respectively, of MRD. Finally, the quality of monoclonal antibody conjugates, as well as the stability and resolution of the flow cytometer, is important. All these factors contribute to, often, large differences in the applicability of specific LAPs for different leukemia cases.

To standardize all the procedures that are needed for LAP definition as well as for quantification of MRD, five centers in Belgium and the Netherlands joined forces in 2004. These centers had ample experience in (at least) four-color flow cytometry. One of these centers had extensive experience with MRD detection in adult AML and served as the reference and coordinating center.

The primary goal of this study was to standardize the identification of LAP at diagnosis. First, we defined a standardized antibody panel and standard operating procedures based on both the MRD experience in the group, the published data and extensive knowledge of the performance of antibody conjugates. Second, we evaluated whether this standardized antibody panel could identify robust LAPs in the vast majority of AML patients. The quality of LAPs was determined on the basis of their presumed specificity, sensitivity and stability. The coordinating center served as reference for the definition of LAPs. Finally, the (improvements in) performance of individual laboratories in identifying LAPs in an initial learning phase and a test phase was evaluated.

MATERIALS AND METHODS

Patients and cells

Sixty-four patients with AML, consecutively presenting during a period of 18 months (July 2004–December 2005) in the participating institutes, were included. The median age was 59 years (range: 9–85; two children of 9 and 15 years were included). FAB (French-American-British) classification distribution was 6 M0, 11 M1, 16 M2, 2 M3, 6 M4, 5 M5, 3 M6, 1 M7, 6 RAEB (3 RAEB and 3 RAEB-t), 1 secondary AML (out of CML), 1 AML with non-Hodgkin's lymphoma and 6 cases with unknown classification. BM of patients with AML, and normal BM from patients with cardiac disease, was obtained after informed consent and according to the institutional protocols. LAPs in normal BM controls were defined as percentages of white blood cells (WBCs).

Participating centers and working plan

The participating centers were required to have the following criteria: (i) longstanding experience in immunophenotyping of leukemia using at least four-color flow cytometry; (ii) access to clinical samples; (iii) experience either in MRD assessment and/or in quantification of low-frequency cell subpopulations; (iv) participation for at least 5 years in external quality control assessment programs for CD34 counting and leukemia/lymphoma immunophenotyping;[17,18] and (v) extensive experience with various antibody–fluorochrome combinations. Five centers from the Netherlands and Belgium participated (for details, see Supplementary Files). Amsterdam (indicated as no. 1) served as the coordinating and reference center, and the other centers were anonymously, in arbitrary sequence, referred to as participants no. 2–5.

Four centers were equipped with a FACScalibur (Becton Dickinson, San Jose, CA, USA) with an argon and red diode laser. CellQuest software was used for analyses of the list mode data (LMD). One participant was equipped with a Beckman Coulter five-color FC500 flow cytometer (Miami, FL, USA) with an argon laser. LMD were analyzed with CXP software (Miami, FL, USA).

The study period was subdivided in two phases: a learning phase with 28 patients and a test phase with 35 patients (for details, see Supplementary Files).

Immunophenotypic labeling

The first goal of the working group was to develop a standard panel of monoclonal antibody combinations to define putative LAPs at diagnosis. This panel was composed of monoclonal antibodies and fluorochromes in four-color combinations based on previous experience with LAP definitions and MRD assessments in AML,[5,19] as well as extensive knowledge of antibody–fluorochrome performance. MoAbs were conjugated with fluorescein isothiocyanate, phycoerythrin (PE), peridinyl chlorophyllin or allophycocyanin for FACSCalibur users, and with fluorescein isothiocyanate, PE, phycoerythrin-Texas Red conjugate (energy coupled dye, ECD) and PE-Cy5 for the Coulter FC500 user (Table 1). The backbone of a LAP consists of CD45, a primitive marker (CD34, CD133 or CD117), a myeloid marker (usually CD13 or CD33) and one of the aberrant markers listed in Table 1. The antibody–fluorochrome combinations were centrally purchased, titrated and diluted to the user concentration, and subsequently distributed to all centers.

LAPs were established in newly diagnosed AML in a two-step approach. In the first step, the panel was used to define the immunophenotype of the blasts in whole BM or peripheral blood at diagnosis. In this so-called *first run*, aberrant expression patterns of individual markers were defined, and, from these, relevant markers for the putative LAPs for MRD assessment (one LAP per tube) were defined. As these markers may originate from different tubes, the finally chosen combinations had to be tested in a so-called *second run*. Because of the large number of LAPs, central distribution of conjugate combinations for the second run was not feasible.

For all relevant monoclonal antibodies, the experience of all centers (advantages, disadvantages and pitfalls) was evaluated in regular meetings of the study group. The final list of antibody–fluorochrome combinations for use in the second run is shown on the website http://www.vumc.nl/afdelingen/hematologie/behandelaars/onderzoek/ under 'clones of monoclonal antibodies'.

Data exchange and data interpretation

When a patient was included in the study in one of the five institutes, the procedure of data exchange was as follows: the LMD of the first run were uploaded by that institute to a website (termed 'X-drive'), with accessibility restricted to the working group members (Figure 1). Using this site, each of the five institutes analyzed these first run LMD files of that particular patient and, independent of each other, defined putative LAPs. These were communicated to the coordinator. From all putative LAPs, the coordinator, according to its experience, defined so-called 'proposed LAPs', fulfilling as many essential criteria as possible. In the quarterly group meetings, these 'proposed LAPs', upon consensus, were termed 'consensus LAPs'. On some occasions, the consensus LAPs deviated from the proposed LAPs. Also, all other putative LAPs were discussed, with performance evaluated for all individual centers.

Table 1. Standard immunophenotypic panel

Tube	*FITC/FITC*[a]	*PE/PE*[a]	*PerCP/ECD*[a]	*APC/PE-CY5*[a]
1	PBS	PBS	CD45	PBS
2	CD34	CD22	CD45	CD117
3	CD15	CD13	CD45	CD14
4	HLA-DR	CD33	CD45	CD11b
5	CD2	CD56	CD45	CD7
6	TdT/CD36	CD133	CD45	CD19

Abbreviations: APC, allophycocyanin; ECD, energy coupled dye; FITC, fluorescein isothiocyanate; PBS, phosphate-buffered saline; PE, phycoerythrin; PerCP, peridinyl chlorophyllin. Before use, monoclonal antibodies were centrally titrated, diluted to working solutions and distributed to the participating institutes. The first tube, containing CD45 in FL3 and PBS in the other channels, was used to discriminate between absence and presence of marker expression. During the test phase of the study, TdT was replaced by CD36. Below is indicated per fluorochrome the antibody clone from BD/antibody clone from Immunotech (markers are in bold for reasons of clarity). In case of other suppliers, these are indicated in between parentheses. FITC: **CD34**: 8G12/581; **CD15**: MMA/MMA; **HLA-DR**: L243/im.357; **CD2**: S5.2/39C1.5; **TdT**: HT-6 (DAKO)/HT-6 (DAKO); and **CD36**: FA6-152/CLB 703 (CLB). PE: **CD22**: S-HCL-1/SJ10.1H11; **CD13**: SJ1D1/SJ1D1; **CD33**: P67.6/D3HL60.251; **CD56**: NCAM16.2/NCAM16.2; and **CD133**: CD133/2 (Miltenyi)/CD133/2 (Miltenyi). PerCP/ECD: **CD45**: 2D1/J33. APC/PE-CY5: **CD117**: 104D2/104D2D1; **CD14**: MoP9/RM052; **CD11b**: D12/Bear 1; **CD7**: M-T701/8H8.1; and **CD19**: SJ25C1/J4.119. [a]Fluorochrome used for BD FCM/fluorochrome used for Beckman Coulter FCM.

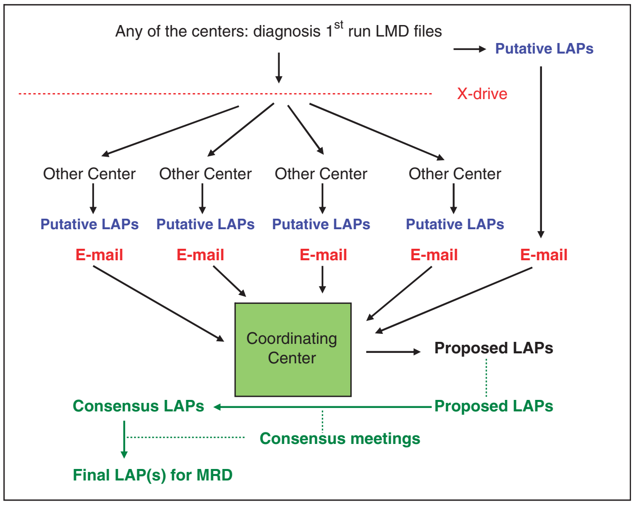

Figure 1. Organization of the study. When any of the five centers entered an AML patient (top of the figure), it ran the monoclonal antibody panel shown in Table 1 and sent the LMD to the other centers. All centers then formulated their putative LAP(s) and sent these to the coordinator, who subsequently designed 'proposed LAP(s)'. Finally, at the quarterly group meetings, 'consensus LAP(s)' were agreed upon, and 'final LAPs' were formulated for subsequent MRD studies.

Statistics

Student's *t*-test was used to evaluate the significance of improvements in performance between the learning and test phases of the study.

RESULTS

Standard immunophenotypic panel (first run)

After the learning phase, tube 6 was modified: TdT was replaced by CD36, as TdT had not contributed to suitable independent LAPs, did not appropriately function as an exclusion marker for M4 and M5 AML types, and, as intracellular marker, hampered a rapid application of the panel.

In the next five sections, the performance of the panel (number and nature of LAPs detected and the sensitivity, specificity and stability of *consensus* LAPs) is described. This part covers the final results after the consensus meetings. Finally, performance of individual centers in both phases is described.

Number and nature of consensus LAPs

In the learning phase, 55 consensus LAPs were found in 28 patients (Tables 2a and b, third column). No LAPs were found in one patient (3.5%). In the test phase, 88 consensus LAPs were found in 35 patients (Tables 2a and b, fourth column). LAPs were found in all patients. For details on numbers of LAPs per patient per study phase, see legends of Table 2. Taken together, one or more LAPs could be identified in almost all AML patients (63/64: 98.5%).

Table 3 shows a detailed overview of consensus LAPs, divided into the well-known four major groups. Important are LAPs that were chosen most frequently as the best (that is, most sensitive and most specific, and stable, as outlined in the next paragraphs). Table 3 shows the leading aberrancies in the first column, with, in the fourth column (printed in boldface), the percentage of patients for whom these aberrancies were chosen as best consensus LAPs. These included cross-lineage expression with CD7 as the leading marker (36% of the patients), asynchronous expression of CD34 and CD133 (20%), asynchronous expression of CD15 combined with a progenitor marker (28%) and absence of HLA-DR (20%) or CD33 (11%) in combination with a progenitor marker. All other aberrancies were considered suitable in $<10\%$ of the patients.

Table 2. Consensus LAPs in terms of sensitivity and specificity

A

Sensitivity[a] (%)	Specificity[a]	Number of consensus LAPs[b,] (% of total, in specific phase)[c]	
		Learning phase	*Test phase*
>50 (high)	High	28 (51)	37 (42)
	Low	7 (13)	7 (8)
20–50 (intermediate)	High	13 (24)	29 (33)
	Low	2 (4)	6 (7)
10–20 (low)	High	4 (7)	6 (7)
	Low	1 (2)	3 (3)
Total		55 (101)	88 (100)

B

Sensitivity[a] (%)	Specificity[a]	Number of best consensus LAPs[b] (% of total, in specific phase)	
		Learning phase	*Test phase*
>50 (high)	High	18 (64)	18 (51)
	Low	2 (7)	3 (9)
20–50 (intermediate)	High	5 (18)	12 (34)
	Low	1 (4)	1 (3)
10–20 (low)	High	1 (4)	1 (3)
	Low	1 (4)	0 (0)
Total		28 (101)	35 (100)

Abbreviation: LAP, leukemia-associated (immuno)phenotype. LAPs are defined by the following preference: first, LAPs with the highest sensitivity ($>50\%$ expression) and best specificity ($<0.1\%$ background) are denoted as 'high/high', and with decreasing preference, this is followed by 'intermediate/high', 'high/low', 'low/high', 'intermediate/low' and, lastly, 'low/low'. [a]LAPs were categorized as a function of sensitivity ('high', $>50\%$ expression; 'intermediate', 20–50% expression; and 'low', 10–20% expression by the leukemic blasts) and specifitity ('high', $<0.1\%$ background; 'low', $\geqslant 0.1\%$ background). [b]In A, total number of consensus LAPs are denoted, with percentages of total number in learning or test phase; in B, total number total number of best consensus LAPs per patient (which thus equals the number of patients) are denoted, with percentages of total number in learning or test phase. [c]In learning phase, no LAPs were found in one patient (3.5%), one LAP in 38%, two LAPs in 41%, three LAPs in 0%, four LAPs in 14% and five LAPs in 3.5% of the cases (not shown in the table). In test phase, LAPs were found in all patients: one LAP in 28%, two LAPs in 26%, three LAPs in 17%, four LAPs in 23% and five LAPs in 6% of the cases (not shown in the table).

In 60 of the 64 patients, at least one primitive marker (CD34, CD117 or CD133; more details about distribution in legends of Table 3) was present in the consensus LAPs. The remaining four cases had LAPs without a primitive marker (three patients), or no LAP at all (one patient).

Sensitivity of the consensus LAPs

LAPs were categorized into the following three groups based on the percentage of LAP-covered leukemic cells (Table 2): (1) $>50\%$ (sensitivity category 'high'); (2) 20–50% ('intermediate'); and (3) 10–20% ('low'). LAPs present on $<10\%$ of blast cells were not considered suitable for MRD monitoring.

Table 2a shows for *all* consensus LAPs that the best sensitivity category ('high') contained 64% (51 + 13%) and 50% (42 + 8%) of the LAPs in the learning phase and test phase, respectively. Taking

Table 3. Summary of consensus LAPs

Aberrant immunophenotypes	*Number of cases*	*% of total number of LAPs*	*Used as best consensus LAPs (% of patients)*[a]
LAPs with immature markers (n = 125)[b]			
Cross-lineage antigen expression (*n = 42*)	*n* = 42	**29.4**	
CD7	26	**18.2**	**36**
CD34^{+} MM^{+} CD7^{+}	11		
CD117^{+}MM^{+}CD7^{+}	8		
CD34^{+}CD117^{+}CD7^{+}	3		
CD117^{+}CD133^{+}CD7^{+}	3		
CD34^{+}CD133^{+}CD7^{+}	1		
CD22	6	4.2	<10
CD34^{+} MM^{+} CD22^{+}	4		
CD117^{+} MM^{+} CD22^{+}	1		
CD34^{+} CD133^{+} CD22^{+}	1		
CD56	5	3.5	<10
CD34^{+} MM^{+} CD56^{+}	4		
CD133^{+} MM^{+} CD56^{+}	1		
CD19	3	2.1	<10
CD117^{+} MM^{+} CD19^{+}	2		
CD34^{+} MM^{+} CD19^{+}	1		
CD2	2	1.4	<10
CD34^{+} MM^{+} CD2^{+}	1		
CD34^{+} CD117^{+} CD2^{+}	1		
Asynchronous antigen expression (*n = 73*)[c]	*n* = 73	**51,0**	
CD34$^{\text{neg}}$ CD133^{+}	14	9.8	**20**
CD34$^{\text{neg}}$ CD133^{+} CD117^{+}	13		
CD34$^{\text{neg}}$ CD133^{++} CD33^{+}	1		
CD15	19	**13.2**	**28**
CD34^{+} MM^{+} CD15^{+}	7		
CD133^{+} MM^{+} CD15^{+}	4		
CD117^{+} MM^{+} CD15^{+}	3		
CD34^{+} CD133^{+} CD15^{+}	3		
CD133^{+} MM^{+} CD15^{+}	1		
CD117^{+} MM^{+} CD15^{+}	1		
CD13$^{\text{neg}}$	9	6.3	<10
CD34^{+} CD117^{+} CD13$^{\text{neg}}$	4		
CD133^{+} MM^{+} CD13$^{\text{neg}}$	2		
CD34^{+} CD133^{+} CD13$^{\text{neg}}$	1		
CD34^{+} MM^{+} CD13$^{\text{neg}}$	1		
CD117^{+} MM^{+} CD13$^{\text{neg}}$	1		
HLA-DR$^{\text{neg}}$	8	5.6	**20**
CD34^{+} CD117^{+} HLA-DR$^{\text{neg}}$	4		
CD34^{+} MM^{+} HLA-DR$^{\text{neg}}$	2		
CD133^{+} CD117^{+} HLA-DR$^{\text{neg}}$	1		
CD117^{+} MM$^{\text{dim}}$ HLA-DR$^{\text{neg}}$	1		
CD33$^{\text{neg}}$	7	4.9	**11**
CD34^{+} MM^{+} CD33$^{\text{neg}}$	7		
CD36^{+} (see legend)	6	4.2	See legends
CD34^{+} CD117^{+} CD36^{+}	3		
CD34^{+} CD133^{+} CD36^{+}	1		
CD133^{+} MM^{+} CD36^{+}	1		
CD117^{+} CD34$^{\text{neg}}$ CD36^{+}	1		
CD11b^{+}	5	3.5	<10
CD117^{+} MM^{+} CD11b^{+}	4		
CD34^{+} MM^{+} CD11b^{+}	1		
CD15$^{\text{neg}}$HLA-DR$^{\text{neg}}$CD117^{+}CD34$^{\text{neg}}$	4	2.8	<10
CD15$^{\text{neg}}$HLA-DR$^{\text{neg}}$CD117^{+}CD34$^{\text{neg}}$	4		
CD45$^{\text{neg}}$	1	0.7	
CD13^{+}CD33^{+}CD45$^{\text{neg}}$CD117^{+}	1		
Overexpression (*n* = 10)	*n* = 10	**7.0**	<10
HLA-DR^{++}	4	2.8	<10
CD117^{+} CD34$^{\text{neg}}$ HLA-DR^{++}	2		
CD34^{+} MM^{+} HLA-DR^{++}	1		
CD117^{+} CD36$^{\text{neg}}$ HLA-DR^{++}	1		
CD13^{++}	4	2.8	<10
CD34^{+} CD133^{+} CD13^{++}	2		
CD34^{+} CD117^{+}CD13^{++}	2		
CD33^{++}	1	0.7	<10
CD34^{+} CD117^{+} CD33^{++}	1		
CD34^{++}	1	0.7	<10
CD117$^{\text{neg}}$ CD19$^{\text{neg}}$CD34^{++}	1		
LAPs without immature markers (n = 18)	*n* = 18	**12.6**	
CD13$^{\text{neg}}$	8	5.6	<10
CD15^{+}CD14$^{\text{neg}}$CD13$^{\text{neg}}$	5		
CD15^{+}CD34$^{\text{neg}}$CD13$^{\text{neg}}$	1		
CD15^{+}CD33^{+}CD13$^{\text{neg}}$	1		
CD11b^{+}HLA-DR^{+}CD13$^{\text{neg}}$	1		
CD56	3	2.1	<10
CD15^{+}CD33^{+}CD56^{+}	1		
CD14$^{\text{neg}}$CD33^{+}CD56^{+}	1		
HLA-DR^{+}CD33^{+}CD56^{+}	1		
CD14$^{\text{neg}}$	2	1.4	<10
CD15^{+}CD13^{+}CD14$^{\text{neg}}$	1		
CD11b^{+}HLA-DR$^{\text{neg}}$CD14$^{\text{neg}}$	1		
HLA-DR$^{\text{neg}}$	2	1.4	<10
CD15$^{\text{neg}}$CD13^{+}HLA-DR$^{\text{neg}}$	1		
CD11b^{+}CD36-DR$^{\text{neg}}$	1		
HLA-DR^{+}	1	0.7	<10
CD15^{+}CD33^{+}HLA-DR^{+}	1		
CD22	1	0.7	<10
CD33^{+}HLA-DR^{+}CD22^{+}	1		
CD33$^{\text{dim}}$	1	0.7	<10
CD34$^{\text{neg}}$CD13^{+}CD33$^{\text{dim}}$	1		
Total[d]	143	100	

Abbreviations: LAP, leukemia-associated (immuno)phenotype; MM, myeloid marker (usually CD13 or CD33, depending on diagnostic phenotype). LAPs identified in the 64 BM samples using the standard immunophenotyping panel. The mean number of LAPs per patients is 2.2 (143 LAPs/64 patients). In only one sample, no LAP could be defined. A primitive marker (CD34, CD117 or CD133) was included in 125 of 143 (87%) of the LAPs. [a]'Best consensus LAP' is defined as the LAP with highest sensitivity and specificity (see Results). More than one LAP may fulfill the same requirement, thereby defining more than one best consensus LAP per patient. As a result in the column used as 'best consensus LAP (% of patients)', the total percentages exceed 100%. [b]CD34 alone was used in 42%, CD117 alone in 27% and CD133 alone in 5%, although any of the four possible *combinations* of at least two primitive markers were used in 18% of the cases (data not shown). In the latter cases, it was often possible to define for the same patient at least two LAPs based on different primitive markers. [c]LAPs with underexpression (usually absence) of markers were almost exclusively present in the group with asynchronous expression (data not shown) and not separately scored. [d]As two or more consensus LAPs were identified in many patients, the total number of consensus LAPs (i.e., 143) exceeded the total number of patients (i.e., 64).

the high and intermediate categories together, these figures were 92% and 90%, respectively.

On performing the same analysis for the *best* consensus LAP (total number is 63, similar to the number of patients with LAP), it was seen that the high category contained in the learning phase 71% of the LAPs and in the test phase 60% of the LAPs. Taking the high and intermediate categories together, these figures were 93% and 97%, respectively (Table 2b).

Specificity of the consensus LAPs

The specificity of a LAP depends on the expression of the aberrant immunophenotype on normal myeloid progenitor and/or mature cells. All participants evaluated the background of the consensus LAPs in normal BM. The median expression on normal myeloid cells usually was $<0.05\%$ (% of WBC), with maximum values generally $<0.1\%$ (Figure 2). This background resulted mainly from aberrant expression on the immature population: LAPs with an immature marker, which represent the vast majority of all consensus LAPs (87%) and of all best consensus LAPs (95%), had up to 10% background aberrancies on the immature population (Figure 2b). CD36 on WBC and on primitive cells was discovered during the study to have higher percentages (Figures 2a and b) and thereafter excluded from further analysis. In LAPs without primitive markers (only 13% of all consensus LAPs and 5% of best consensus LAPs), median LAP expression on normal BM WBC was usually $<0.1\%$ (data not shown).

The specificity of a consensus LAP was now categorized as (1) 'high' ($<0.1\%$ background) or (2) 'low' ($\geqslant 0.1\%$ background). The two specificity categories are presented in Table 2a and distributed over the three sensitivity categories. The high-specificity category contained in the learning phase 82% (51 + 24 + 7%) and in the test phase 82% (42 + 33 + 7%) of all consensus LAPs.

On performing the same analysis for the *best* consensus LAP (total number 63), it was seen that the highest specificity category represented 86% of the patients in the learning phase and 88% in the test phase (Table 2b).

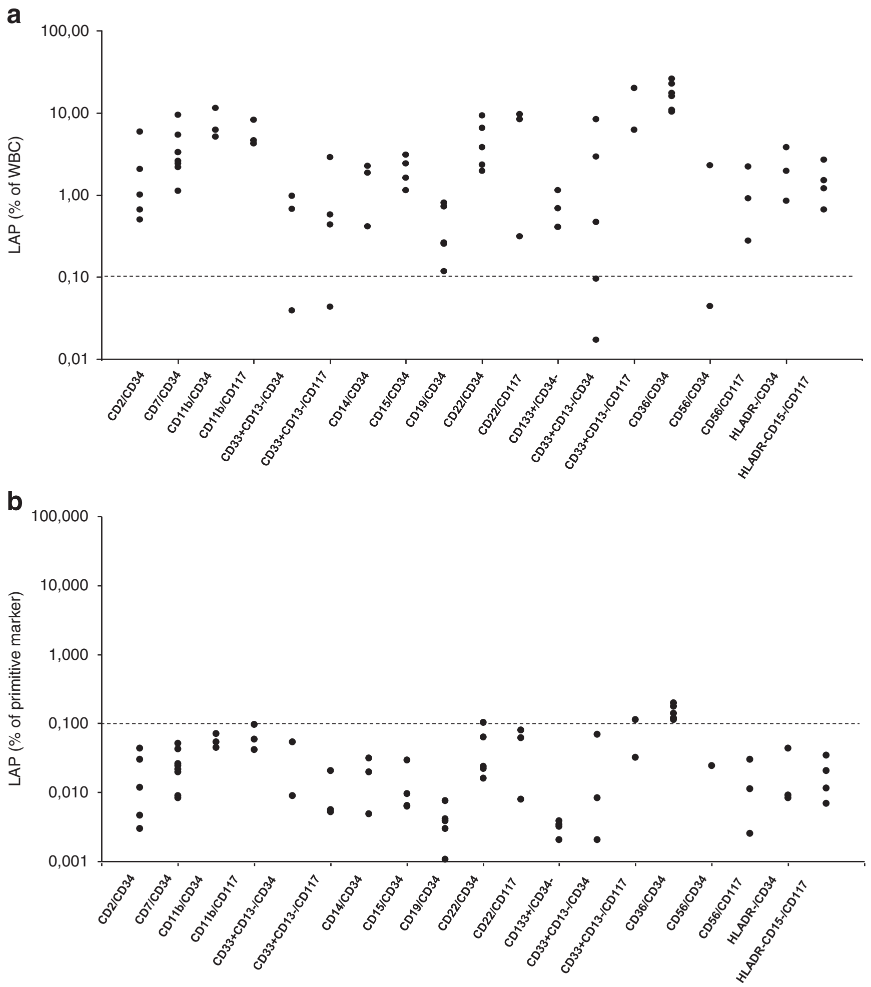

Figure 2. Aberrant expression of marker/marker combinations on WBC and immature cell fractions in normal BM. Different LAP immunophenotypes tested in normal BM. All LAPs contained CD45 and, if not indicated in the figure, a myeloid marker (CD13 or CD33) or CD117. (**a**) LAP expression as % of WBC. Note the relatively high expression of CD34$^+$CD36$^+$. (**b**) Aberrant marker expression as a percentage of primitive marker compartment (CD34 or CD117). Note logarithmic scale in **a** and **b**.

Combined analysis of sensitivity and specificity

Ultimately, the quality of the consensus LAP has to be acquired on the *combined evaluation of sensitivity and specificity*. The best LAP possesses the highest sensitivity ($>$50% expression) and the best specificity ($<$0.1% background). The order of quality of other LAP categories is presented in the legends of Table 2. Assuming that the high- and intermediate-sensitivity categories, together with high specificity, will offer the best consensus LAPs (63 in total), Table 2b shows that in 82% (64 + 18%) of cases in the learning phase and in 85% (51 + 34%) in the test phase, LAPs met this demand (further details are in Supplementary Files). It can thus be concluded that, in the majority of the 63 evaluable patients, relatively high-quality ($>$20% expression and $<$0.1% background reactivity) LAPs can be designed.

Stability of LAPs

Scoring for a good LAP based on stability did not include a learning and test phase, but it was performed using a list of expression of markers in paired diagnosis–relapse samples. We have performed an extensive comparison between diagnosis and relapse in 33 AML patients (18 different antigens in 125 diagnosis–relapse comparisons; Supplementary Table 1). In general, LAPs are relatively stable in most cases with a notable exception for CD19. For details, see Supplementary Results and Supplementary Table 1.

Performance of individual centers

The data discussed above constitute the final outcome of all consensus meetings. Next, we describe the performance of individual centers.

Missed LAPs. In the learning phase, centers 2–5 missed 39–63% of the consensus LAPs, whereas the coordinator (no. 1) missed 11% (Table 4). In the test phase, these percentages dropped to 27–40% and 3%, respectively (Table 4). Thus, the improvement was 12–33% for centers 2–5 and 8% for the coordinator. Supplementary Figure 2 shows a detailed overview of the numbers of missed LAPs per institute per phase.

Number of patients with no second LAP defined; comparison between the institutes. We then studied how many *patients* would not have been eligible for follow-up evaluation because of the missing of all LAPs present. For a patient to be eligible for MRD evaluation, at least one consensus LAP should have been defined. The answer, based on the data in Supplementary Figure 2, has been summarized in Figure 3a: in the learning phase, the percentages of patients apparently ineligible for MRD detection ranged from 21 to 33% for centers 2–5, whereas the coordinator identified at least one LAP in all. In the test phase, results clearly improved: the percentages of patients without any identified LAP ranged from 7 to 18%, whereas the coordinator had no misses.

The best policy to avoid missing MRD due to immunophenotypic shifts may be to define, whenever possible, at least two different LAPs. The performance of all centers to define a second LAP in patients where such LAP had indeed been formulated at the consensus meetings is shown in Figure 3b. In the learning phase, such a second LAP was missed in 8–29% of patients by centers 2–5 and in 10% of patients by the coordinator. These percentages were slightly lower in the test phase: a second LAP was missed in 9–20% of patients by centers 2–5 and in 6% by the coordinator. Thus, in addition to patients in which all LAPs had been missed (Figure 3a), a second LAP may be missed in some instances, where a first LAP had been successfully identified (Figure 3b).

Types of missed LAPs. In a next analysis, we investigated which types of LAPs are missed. LAPs missed by more than one center in relatively high percentages were those with asynchronous antigen expression (31–48%), with the exception of LAPs with an absence of marker(s) (for example, $CD133^+CD34^-$, $CD15^-HLA\text{-}DR^-$, $CD33^-$ and $HLA\text{-}DR^-$; see Table 3), LAPs with marker overexpression (25–88%) and LAPs without immature markers (14–50%) (Supplementary Table 2). The identification of LAPs with cross-lineage marker expression was generally most successful (7–32% misses). Performance of the individual LAPs is shown in Supplementary Table 3. Although, upon this further subdivision, numbers become too low to reliably evaluate, institutes were observed to perform well for the cross-lineage LAP with CD7, and also for LAPs with aberrant absence of marker expression (see above).

Irrespective of the type of LAP, it may be argued that the sensitivity and specificity category may also affect the success of identification of a LAP. Figure 4 shows the results for sensitivity only, because the specificity categories harbored too low numbers for reliable conclusions. The main conclusion (details in Supplementary Results) is that the centers performed much better for high-sensitivity category compared with intermediate and low-sensitivity category. However, substantial improvement was seen from the learning phase to the test phase, ending with almost similar performance in the high and intermediate category in the test phase.

Incorrectly assigned LAPs. Although missed LAPs may prevent evaluation of a patient's MRD, incorrect assignment of a LAP may result in unreliable assessments of MRD cell frequency. Incorrectly designed LAPs mainly resulted from underestimation of background LAP expression in normal BM. In the learning phase,

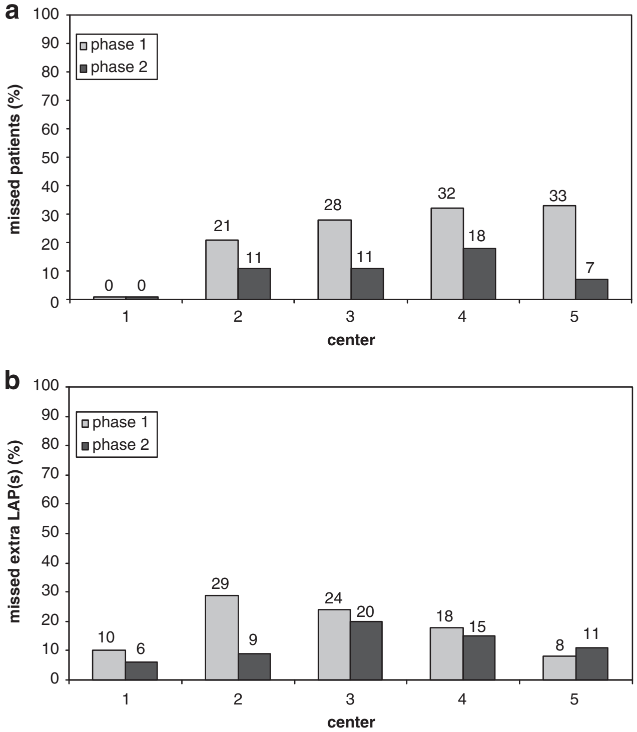

Figure 3. Missed consensus LAPs. For 63 of the 64 patients reported here, at least one consensus LAP could be agreed upon. Cases with no LAP defined for a particular institute and a particular patient were scored as 'missed patient'. Institutes 1–5 analyzed 29, 24, 29, 28 and 12 samples in the learning phase (gray bars) and 35, 35, 35, 34 and 27 samples in the test phase (black bars), respectively. Panel **a** shows the proportions of patients in whom no LAPS were defined and panel **b** shows the proportions of patients in whom extra LAPs were missed. (**a**) For institutes 1–5, percentages of patients who would not have been eligible for MRD assessment are shown (no LAPs were defined, whereas at least one consensus LAP was present). Improvement was 0%, 10%, 17%, 14% and 26% for institutes 1–5, respectively. Differences between the two phases were significant: $P = 0.035$ (all institutes: $n = 5$) and $P = 0.016$ (institutes 2–5). (**b**) For institutes 1–5, the percentages of patients in whom only one LAP was defined, whereas at least two consensus LAPs were present, are shown. Only institute 2 had made considerable improvement (20%). Note that this figure does not include learning effects for all defined LAPs: these have been shown earlier in Table 4. In that table, all institutes made progression, which apparently is merely due to increased performance in defining LAPs *additional to* the two defined for Figure 3b.

Table 4. Percentages of missed LAPs by the different centers in learning and test phase

Phase	*Institute 1*		*Institute 2*		*Institute 3*		*Institute 4*		*Institute 5*	
	Missed	*%*	*Missed*	*%*	*Missed*	*%*	*Missed*	*%*	*Missed*	*%*
1	6/55	11	18/46	39	28/55	51	34/54	63	20/33	61
2	3/88	3	24/88	27	32/88	36	34/84	40	18/65	28
Δ		− 8		− 12		− 15		− 23		− 33

Abbreviation: LAP, leukemia-associated (immuno)phenotype. A LAP was noted as 'missed' in cases where a LAP had not been defined with the aberrant marker agreed upon in the consensus LAP. Only institutes 1 and 3 evaluated all 64 samples, with a maximally reachable number of 143 LAPs. Institutes 2, 4 and 5 evaluated 59, 62 and 39 of the 64 samples, respectively. In the samples evaluated by these institutes, the maximally reachable number of LAPs was 134, 138 and 98, respectively, and percentages of missed LAPs were calculated using the latter numbers. 'Δ' indicates difference in % between learning phase (phase 1) and test phase (phase 2). Differences between the two phases were significant: $P = 0.015$ (all institutes) and $P = 0.022$ for institutes 2–5.

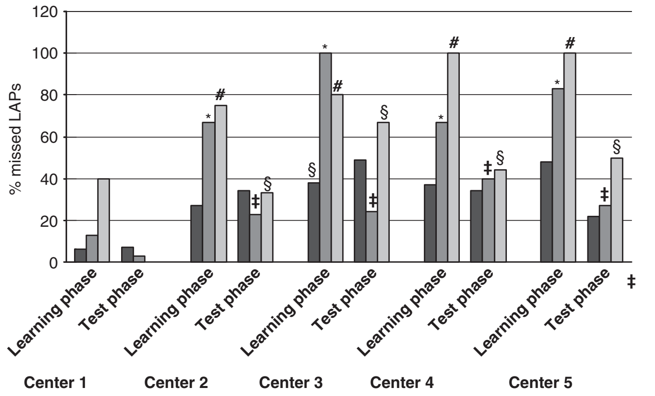

Figure 4. Performance of individual institutes in different sensitivity categories in the learning and test phase. The figure shows the performance of the four centers to define LAPs, now subdivided into the three sensitivity categories (i.e., high (>50%), black; intermediate (20–50%), dark gray; and low (10–20%), light gray). Missed LAPS in a sensitivity category are expressed as percentage of the maximal number of consensus LAPs defined in that specific sensitivity category. In the learning phase, the percentage of missed LAPs was significantly higher in the intermediate-sensitivity (marked $*P=0.01$) and in the low-sensitivity category (marked $^{\#}P=0.001$) than in the high-sensitivity category (all centers included). These differences were also significant when the coordinating center (1) was excluded from analysis (*P*-values ranging between 0.01 and 0.05). When comparing the learning and test phases, for centers 2–5 there was an improvement in the intermediate-sensitivity category (marked $^{\ddagger}P=0.016$). There was also improvement in the lowest sensitivity category (marked $^{\S}P=0.024$).

centers performed experiments on background staining in normal BM and evaluated specificity. Specificity of LAPs was an important issue at the consensus meetings. Incorrectly designed LAPs were only scored as such during the test phase and are shown (per sample and per institute) in Supplementary Figure 2 (green symbols). The percentages of incorrectly designed LAPs were low in institutes 1 and 2, but considerably higher in institutes 3 and 4. The number of LAPs was too low for reliable conclusions for institute 5. Such incorrect designs included the omission of exclusion markers, incorrect interpretation of levels of expression and definition of incorrect LAPs (for detailed information, see Supplementary Results). No specific type of incorrectly designed LAPs, common to all participants, could be defined, except for the fact that, for LAPs defined by the absence of expression of markers such as HLA-DR, inclusion of an exclusion marker is recommended.

DISCUSSION

The current study was undertaken to establish criteria necessary to enable assessment of MRD in clinical studies in a multicenter setting. Such studies should establish whether or not MRD has the prognostic value reported by several single-institute retrospective studies, and at what time points and for which specific patient groups MRD assessment can be a part of clinical decision making. Such approach has in the meanwhile been applied in childhood AML, although in a single-center approach.[12]

When defining criteria to be fulfilled for a multicenter approach, we defined two distinct studies: first, definition of LAP at diagnosis and, second, the use of such LAPs to detect and quantify MRD during follow-up. The present study deals with the diagnostic part. Altogether, the final success rate of defining at least one consensus LAP was high (only 1 out of 64 samples had no LAPs at all), whereas in two-third of the samples two or more independent consensus LAPs could be defined. However, this result was achieved after multiple analyses and extensive group discussions and was based for a large part on the experience of the coordinating center. The complexity to define LAPs by relatively unexperienced centers is evidenced by the high percentages of missed LAPs by centers 2–5, even in the second, testing phase. Some cross-lineage aberrancies, as well as those asynchronous aberrancies that are characterized by complete lack of a marker, were apparently the easiest to define. Other asynchronous aberrancies, antigen overexpression and aberrancies on mature cells were more difficult to establish. In addition, in the initial learning phase, the success rate decreased with decreasing sensitivity (that is, represented by blast coverage percentage of LAPs).

With time, a learning effect, evidenced by the success rate of defining LAPs, was seen. This result was partly because of the thorough discussions at the group meetings. Ultimately, in the test phase this resulted in a success rate (that is, at least one LAP defined) of 82–93% for the relatively inexperienced institutes 2–5 in the 35 evaluable samples. Improvement was especially evident in a category with intermediate sensitivity; in the test phase, LAPs were identified with equal success rate in the high-sensitivity and the intermediate-sensitivity category. However, the overall results implicate that, because of missed LAPs, 7–18% of the patients would not have been evaluable for MRD according to these centers.

Furthermore, in another 9–20% of cases, a center defined only a single LAP, whereas at least two LAPs were finally defined in the consensus meetings. These patients would be more likely to become false negative for MRD if immunophenotypic shifts would have occurred. It has been reported by others and by us[5,11,16] that such shifts occur frequently, although it has not always led to the disappearance of markers. In case only one aberrant marker is present, it is advised to design a second staining in which the nonaberrant markers that are included in the LAP are replaced by other, nonaberrant, markers. In this way, the risk of loss of these normal markers, which could also result in false negativity, is minimized. For example, a 'primitive' marker, CD133, may occasionally become negative; replacement by, for example, CD34 would then be adequate. This condition was met in the current study because in 18% of cases at least two primitive markers among CD34, CD133 and CD117 were present. In most cases, it is possible to replace a myeloid marker by another. As the expression of CD33, CD13 and CD117 is quite stable, the design of the same LAPs with alternative myeloid markers is not urgent.

Our group had agreed that LAPs should cover at least 10% of the blast cells at diagnosis: first, because 'background' levels in normal BM may amount to this percentage and, second, because lower frequencies will significantly hamper the sensitivity necessary for MRD detection.

Another important factor for the quality of a LAP is specificity, which is determined by LAP expression on control normal or regenerating BM. Cell populations with the immunophenotype of LAPs are present in frequencies varying from 0.001 to 0.1% in control BM. In practice, the contribution of these 'LAP-like' events to specific AML-derived, LAP + cells will have to be defined in the setting of MRD. This interaction will strongly depend on the level and intensity of expression of the aberrant markers on the normal versus AML cell populations. AML cells at diagnosis may occupy the so-called 'empty spaces'—that is, positions of antigen expression absent in normal cells; however, because of even moderate immunophenotypic changes during or after therapy, AML-specific LAP + events may interfere with normal LAP + events. Even if stringent precautions have been taken to standardize antibody–fluorochrome combinations, as well as instrumental setup and calibration over time, there is no guarantee that the AML LAP + population will not be 'contaminated' by normal events with 'LAP-like' immunophenotypes. The impact of 'background LAP' on the reliable detection of specific (AML) LAP + events is considerable: up to 0.1% of WBC and up to 10% of primitive marker compartment is possible; as a result, specific detection of MRD $<0.1\%$ is compromised (Figure 2). On the basis of its very high background levels, at the end of the learning phase, CD36 was concluded not to represent a valuable marker for MRD.

The type of LAP and their distribution over patient groups differ within the literature. In the present study, the frequency of LAPs with asynchronous expression was twice as high as for cross-lineage aberrancies, which is in agreement with some other adult studies performed with these two types of major aberrancies.[14] However, in two previous studies on adult AML[5] and childhood AML[11] in which we were involved, the frequencies of both LAP types were more or less equal.

Our study also indicated that MRD studies may be performed using different instruments: in our case, four Becton Dickinson and one Beckman Coulter (Miami, FL, USA) flow cytometer were used. This is of importance in view of international clinical studies in which MRD assessment may be dealt with by (core) centers that may use different brands of flow cytometers.

The large differences in specificity, as well as LAP coverage on blasts, represent a problem when cutoff levels have to be defined for clinical decision making: although in the case with high specificity and high sensitivity, MRD levels of 0.01% (that is, 1:10 000) can be accurately measured, in other cases levels of only 0.1% or even lower may be the maximally reachable sensitivity. Improvements can be expected from the use of more colors, as shown by Voskova *et al.*[13], in the transition from four to five colors. The use of at least eight-color flow cytometry, now feasible in many institutes, will further improve flow-cytometric MRD analysis. The use of novel software tools (for example, Infinicyt (Cytognos, Salamanca, Spain)) for defining LAPs in an objective manner may further improve flow-cytometric MRD analysis.

In conclusion, our study has shown that immunophenotypic MRD assessment is a complex process that requires specific experience that is not covered by a large experience in only standard diagnostic immunophenotyping. The quality of MRD assessment can only be guaranteed after a substantial period of training under the guidance of highly experienced center(s). Therefore, we recommend that, currently, the number of centers that participate in large-scale multicenter MRD studies be restricted to a few core centers with MRD experience.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The European Society for Clinical Cell Analysis (ESCCA) supported the study with a grant from BD Biosciences (CA, USA) for the screening panel of monoclonal antibodies for the FACSCalibur users. We thank Ellen van Lochem and Jaco Kraan for their contributions during the initial phase of the study and Jennita Slomp (Medisch Spectrum Twente, The Netherlands) for critically reading the manuscript.

REFERENCES

1 Béné MC, Kaeda JS. How and why minimal residual disease studies are necessary in leukemia: a review from WP10 and WP12 of the European LeukaemiaNet. *Haematologica* 2009; **94**: 1135–1150.

2 San Miguel JF, Vidriales MB, Lopez-Berges C, az-Mediavilla J, Gutierrez N, Canizo C *et al.* Early immunophenotypical evaluation of minimal residual disease in acute myeloid leukemia identifies different patient risk groups and may contribute to postinduction treatment stratification. *Blood* 2001; **98**: 1746–1751.

3 San Miguel JF, Martínez A, Macedo A, Vidriales MB, López-Berges C, González M *et al.* Immunophenotyping investigation of minimal residual disease is a useful approach for predicting relapse in acute myeloid leukemia patients. *Blood* 1997; **90**: 2465–2470.

4 Venditti A, Buccisano F, Del Poeta G, Maurillo L, Tamburini A, Cox C *et al.* Level of minimal residual disease after consolidation therapy predicts outcome in acute myeloid leukemia. *Blood* 2000; **96**: 3948–3952.

5 Feller N, van der Pol MA, van Stijn A, Weijers GW, Westra AH, Evertse BW *et al.* MRD parameters using immunophenotypic detection methods are highly reliable in predicting survival in acute myeloid leukaemia. *Leukemia* 2004; **18**: 1380–1390.

6 Kern W, Voskova D, Schoch C, Hiddemann W, Schnittger S, Haferlach T. Determination of relapse risk based on assessment of minimal residual disease during complete remission by multiparameter flow cytometry in unselected patients with acute myeloid leukemia. *Blood* 2004; **104**: 3078–3085.

7 Sievers EL, Lange BJ, Alonzo TA, Gerbing RB, Bernstein ID, Smith FO *et al.* Immunophenotypic evidence of leukemia after induction therapy predicts relapse: results from a prospective Children's Cancer Group study of 252 patients with acute myeloid leukemia. *Blood* 2003; **101**: 3398–3406.

8 Campana D, Coustan-Smith E. Detection of minimal residual disease in acute leukemia by flow cytometry. *Cytometry* 1999; **38**: 139–152.

9 Langebrake C, Creutzig U, Dworzak M, Hrusak O, Mejstrikova E, Briesinger F *et al.* Residual disease monitoring in childhood acute myeloid leukemia by multiparameter flow cytometry: the MRD-AML-BFM Study Group. *J Clin Oncol* 2006; **24**: 3686–3692.

10 Buccisano F, Maurillo L, Del Principe MI, Del Poeta G, Sconocchia G, Lo-Coco F *et al.* Prognostic and therapeutic implications of minimal residual disease detection in acute myeloid leukemia. *Blood* 2012; **119**: 332–341.

11 van der Velden VH, van der Sluijs-Geling A, Gibson BE, te Marvelde JG, Hoogeveen PG, Hop WC *et al.* Clinical significance of flowcytometric minimal residual disease detection in pediatric acute myeloid leukemia patients treated according to the DCOG ANLL97/MRC AML12 protocol. *Leukemia* 2010; **24**: 1599–1606.

12 Rubnitz JE, Inaba H, Dahl G, Ribeiro RC, Bowman WP, Taub J *et al.* Minimal residual disease-directed therapy for childhood acute myeloid leukaemia: results of the AML02 multicentre trial. *Lancet Oncol* 2010; **11**: 543–552.

13 Voskova D, Schnittger S, Schoch C, Haferlach T, Kern W. Use of five-color staining improves the sensitivity of multiparameter flow cytometric assessment of minimal residual disease in patients with acute myeloid leukemia. *Leuk Lymphoma* 2007; **48**: 80–88.

14 Al-Mawali A, Gillis D, Lewis I. The role of multiparameter flow cytometry for detection of minimal residual disease in acute myeloid leukemia. *Am J Clin Pathol* 2009; **131**: 16–26.

15 Macedo A, San Miguel JF, Vidriales MB, López-Berges MC, García-Marcos MA, Gonzalez M *et al.* Phenotypic changes in acute myeloid leukaemia: implications in the detection of minimal residual disease. *J Clin Pathol* 1996; **49**: 15–18.

16 Baer MR, Stewart CC, Dodge RK, Leget G, Sulé N, Mrózek K *et al.* High frequency of immunophenotype changes in acute myeloid leukemia at relapse: implications for residual disease detection (Cancer and Leukemia Group B Study 8361). *Blood* 2001; **97**: 3574–3580.

17 Van Wering ER, van Lochem EG, Leenders M, van der Sluijs-Gelling AJ, Wind H, Gratama JW *et al.* Three-color flowcytometric analysis of mature and immature hematological malignancies. A guideline of the Dutch Foundation for Immunophenotyping of Hematological Malignancies (SIHON). *Biol Regul Homeost Agents* 2004; **18**: 313–326.

18 Levering WH, Preijers FW, van Wieringen WN, Kraan J, van Beers WA, Sintnicolaas K *et al.* Flow cytometric CD34 + stem cell enumeration: lessons from nine years external quality assessment within the Benelux countries. *Cytometry B Clin Cytom* 2007; **72**: 178–188.

19 Feller N, Schuurhuis GJ, van der Pol MA, Westra G, Weijers GW, van Stijn A *et al.* High percentage of CD34-positive cells in autologous AML peripheral blood stem cell products reflects inadequate in vivo purging and low chemotherapeutic toxicity in a subgroup of patients with poor clinical outcome. *Leukemia* 2003; **17**: 68–75.

Supplementary Information accompanies this paper on Blood Cancer Journal website (http://www.nature.com/bcj)