Leukemia Research 76 (2019) 39–47

Contents lists available at ScienceDirect

Leukemia Research

journal homepage: www.elsevier.com/locate/leukres

Research paper

Immunophenotypic measurable residual disease (MRD) in acute myeloid leukemia: Is multicentric MRD assessment feasible?

Rik A. Brooimans[a,b,*], Vincent H.J. van der Velden[a], Nancy Boeckx[c], Jennita Slomp[d], Frank Preijers[e], Jeroen G. te Marvelde[a], Ngoc M. Van[b], Antoinette Heijs[d], Erik Huys[e], Bronno van der Holt[f], Georgine E. de Greef[f], Angele Kelder[g], Gerrit Jan Schuurhuis[g]

[a] Department of Immunology, Laboratory Medical Immunology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[b] Laboratory of Clinical and Tumor Immunology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[c] Laboratory of Experimental Transplantation, University of Leuven, Leuven, Belgium
[d] Department of Clinical Chemistry, Medisch Spectrum Twente/Medlon, Enschede, The Netherlands
[e] Department of Laboratory Medicine-Laboratory for Hematology, Radboud UMC, Nijmegen, The Netherlands
[f] Department of Hematology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[g] Department of Hematology, VU University Medical Center, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
AML
Measurable residual disease
Flow cytometry
Multicentre

ABSTRACT

Flow-cytometric detection of now termed measurable residual disease (MRD) in acute myeloid leukemia (AML) has proven to have an independent prognostic impact. In a previous multicenter study we developed protocols to accurately define leukemia-associated immunophenotypes (LAIPs) at diagnosis. It has, however, not been demonstrated whether the use of the defined LAIPs in the same multicenter setting results in a high concordance between centers in MRD assessment. In the present paper we evaluated whether interpretation of list-mode data (LMD) files, obtained from MRD assessment of previously determined LAIPs during and after treatment, could reliably be performed in a multicenter setting. The percentage of MRD positive cells was simultaneously determined in totally 173 LMD files from 77 AML patients by six participating centers. The quantitative concordance between the six participating centers was meanly 84%, with slight variation of 75%–89%. In addition our data showed that the type and number of LAIPs were of influence on the performance outcome. The highest concordance was observed for LAIPs with cross-lineage expression, followed by LAIPs with an asynchronous antigen expression. Our results imply that immunophenotypic MRD assessment in AML will only be feasible when fully standardized methods are used for reliable multicenter assessment.

1. Introduction

Acute myeloid leukemia (AML) is a heterogeneous malignant disease characterized by the accumulation of immature myeloid progenitor cells, which leads to anemia, thrombocytopenia and impaired immunity. Treatment with intensive chemotherapy regimens of adult AML patients who are 60 years of age or younger results in hematologic remission in about 70–90% of patients, but at least 30% of these patients will experience a relapse [1]. Remaining cells in the bone marrow after chemotherapy treatment are thought to be responsible for the relapse. The small number of malignant cells was previously termed minimal residual disease, and is presently referred to as "measurable residual disease" (MRD) [2]. Early relapse prediction is therefore of high importance, so that post-remission therapy can be applied on time and to decide on transplant or not. Treatment response assessment by the quantification of MRD is based on specific morphologic [3], immunophenotypic [4] or genotypic [5,6] aberrancies, and is one of the predictors of relapse in AML. Often the residual leukemic cells are present at levels below the sensitivity of conventional microscopic examination and the detection therefore requires more sensitive methods. Real time quantitative PCR is the most sensitive method for the detection of MRD in AML, but molecular markers suitable for follow-up are available in about 60% of AML patients only [7]. Immunophenotypic detection of MRD by multiparameter flow cytometry (MFC) provides an useful alternative in the detection of MRD because it allows: simultaneous identification and quantification of tumors cells at

[*] Corresponding author at: Laboratory Medical Immunology, Department of Immunology, Erasmus MC, University Medical Center Rotterdam, Wytemaweg 80, 3015, CN, Rotterdam, The Netherlands.
E-mail address: r.brooimans@erasmusmc.nl (R.A. Brooimans).

https://doi.org/10.1016/j.leukres.2018.11.014
Received 14 August 2018; Received in revised form 1 November 2018; Accepted 26 November 2018
Available online 27 November 2018

the single-cell-level, while fast and relatively cheap evaluation of high cell numbers makes it possible to reliably define aberrant cell surface antigen expression on AML blasts. These aberrant antigen combinations of the leukemic cells, known as Leukemia-Associated Immunophenotypes (LAIPs), can be identified in nearly all patients with AML at time of diagnosis [8] and are not, or at very low frequencies, detectable on normal blasts. Immunophenotypic detection of MRD in AML assessed after different therapies has proven to deliver independent prognostic impact, mainly in single-institute studies [9–14], but recently also in prospective multicentre studies [15,16]. Moreover, measuring the extent of leukemic clearance by MRD assays determines the resistance to therapy and thereby, importantly, can assess the relapse risk in an individual patient independently from other risk factors. As a result, the early detection of a forthcoming relapse might result in risk adapted therapies.

For a previous multicenter study a working group was founded of Dutch and Belgian laboratories with ample experience in flow cytometry that developed common protocols to accurately define LAIPs at diagnosis required to establish MRD during/after treatment. In that study we have shown that immunophenotypic MRD assessment is a complex process that requires specific experience and standardization in identification of LAIPs between the laboratories involved [17]. It has, however, not been demonstrated whether the use of the defined LAIPs in the same multicenter setting results in a high agreement between centers in MRD assessment in follow up samples of AML treated patients.

The aim of the present study was to determine whether interpretation of list-mode data (LMD) files obtained from immunophenotypic MRD assessment and based on LAIPs at diagnosis, can reliably be evaluated for MRD detection, so that multiple laboratories could concertedly provide MRD information of similar quality in multicenter trials.

2. Materials and methods

2.1. Patients and samples

Seventy seven 77 patients suffering from AML, consecutively presenting and treated according to HOVON clinical trial 42a (www.hovon.nl) in a period of about 3 years in the participating institutes, were included. All patients had a cytopathologically confirmed diagnosis according to the WHO classification (excluding acute promyelocytic leukemia). Bone marrow of patients treated for AML was obtained at diagnosis and during follow-up after first and second cycles of chemotherapy according to the institutional protocols with given general informed consent with central approval number Erasmus MC, METC 2000-220.

2.2. Participating laboratories/working group

All six participating institutes (arbitrary referred to as center 1–6) are part of the AML-MRD flow cytometry working group of the Dutch Cytometry Society that was installed in 2004 with the main objective to define the prerequisites for reliable multicenter MRD assessment in studies with patients treated for AML. At that time all the laboratories had longstanding experience in immunophenotyping of leukemia using at least 4-color flow cytometry and had experience in MRD assessment.

2.3. Data exchange and data reporting

All LMD files of the LAIPs run by the participating institutes on the BM samples at diagnosis were already available in the coordinating center (VUMC, Amsterdam) [17]. These LAIPs were also actually designed and used by the institutes for MRD assessment of their own AML patients and uploaded to the coordinating center for central MRD analysis.

At twelve different occasions during the three year study period, the coordinating center selected LMD files of 5–7 cases for evaluation of multicentric MRD assessment. Each case consisted of LMD files of the selected LAIPs at diagnosis and the acquired data at follow-up. The procedure of data exchange to the individual centers has been previously described [17]. Each institute analyzed the provided LMD files for MRD assessment of the acquired follow-up samples of that particular patient, using the same gating strategy according previously established guidelines within the working group [11,16,17]. In short, analysis of LAIP-positive cells included multiple backgating steps to ensure that, compared with diagnosis, the LAIP-positive cells show fairly identical positions in forward scatter channel/side scatter channel and CD45 expression. Each center reported independently (I) the number of LAIP positive cells as percentage within the WBC compartment (%LAIP), and (II) the interpretation of these analyzed LAIP data, by notifying whether MRD is detectable and, if present, the percentage of MRD. Based on our experience and outcome in previous studies [11,16], the working group had agreed that the cut-off for MRD positivity in the present study should also be set at 0.1%. Therefore all data reported between 0.05 and 0.1% were classified as MRD negative in the present analysis, whereas all LAIP percentages reported below 0.05% were considered as background values and therefore converted in the calculations to negative or zero. Concordance was determined by a qualitative and a quantitative analysis. In case of qualitative analyses, the reported % LAIP and MRD by the individual centers was classified as positive when the percentage was 0.1 or higher and negative when the percentage was below 0.1. The majority votes (MRD positive or MRD negative) per sample were used, as denominators for comparisons. In those cases in which votes on both sides were equal, the vote of the coordinator was used to finally define the expected result. Concordance regarding quantitative analyses of percentage LAIP and MRD was assessed by defining the median value of all the individual reported percentages of each case as expected value.

In quarterly meetings of the working group the LAIP and MRD percentages were discussed, and the performance of the participating institutes were evaluated. While the coordinating center had access to all files, the following precautions were taken to circumvent bias: MRD assessment results from the patients uploaded to the coordinating center by the five other centers, were downloaded by the coordinating center only after having produced their own MRD analysis of each uploaded sample. To avoid bias in cases where this was not possible, the technician who performed the actual MRD analysis in the coordinating center was different from the one who gathered the X-drive data from the other five centers.

2.4. Analyses of MRD

LMD files were exchanged with delivery of the identity of the LAIPs as established at diagnosis, as well as the corresponding original diagnostic LMD files. MRD analysis was performed as previously described [11]. Gating strategy of MRD in BM after the 1st or 2nd cycle of chemotherapy is based on the LAIP expression in diagnosis AML material. Analysis of LAIP-positive cells included multiple backgating steps to ensure that, compared with the immunophenotyping at diagnosis, the LAIP-positive cells show fairly identical positions in forward scatter channel/side scatter channel and CD45 expression. By using this method, LAIP-positive cell populations could be distinguished from background expression in the gate. MRD percentage was defined as the percentage of LAIP-positive cells within the WBC compartment.

2.5. Statistics

Concordance of independent MRD data was assessed with the Fischer's exact test.

The intra-class correlation coefficient (ICC) was used to estimate correlations of MRD values between individual measurements and

Table 1
Qualitative [a] concordance in classification of LAIP-positive cells by the different centers.

Center	1		2		3		4		5		6		All		Range
exp,obs[b]	#	%	#	%	#	%	#	%	#	%	#	%	#	%	%
pos,pos	63	36	61	38	56	34	67	40	46	35	39	36	332	37	
neg,neg	89	51	80	49	69	42	84	50	65	50	59	55	446	49	
concordance		87		87		76		90		85		91		86	76–91
neg,pos[c]	11	11	16	17	28	29	8	9	10	13	1	2	74	14	
pos,neg[d]	10	14	5	8	13	19	10	13	10	18	8	17	56	14	
n =	173		162		166		169		131		107		908		
sensitivity[e]		86		92		81		87		82		83		86	81–92
specificity[f]		89		83		71		91		87		98		86	71–98

[a] all percentages of LAIP equal or above cutoff (≥0.1%) are classified as 'positive', below (< 0.1%) as 'negative'.
[b] see method section for definition of expected versus observed result.
[c] false-positives = expected result is negative, whereas observation is positive.
[d] false-negatives = expected result is positive, whereas observation is negative.
[e] [(pos,pos)/[(pos,pos) + (pos,neg)]] * 100.
[f] [(neg,neg)/[(neg,pos) + (neg,neg)]] * 100.

between average measurements made on the same target. The ICC was calculated according to Shrout and Fleiss (Model 2, i.e., ICC2.1) [18]. The cut-off point of 0.75 ICC discriminates between good and moderate to poor agreement of observed versus expected results.

3. Results

3.1. Analysis of LAIP-positive cells

In the present study six participating centers indepentdently analysed a total of 173 MRD LMD files from 77 patients (median: 2.2 per patient, range: 1–4) according to analysis strategy of the working group and reported the percentage LAIP positive cells of the total leucocytes to the coordinating center. Regarding the qualitative concordance, the expected nominated positives and negatives were compared with the observed classified positives and negatives (see Methods). Using the cut-off of ≥ 0.1% for a positive sample, analysis of the LMD files resulted in 73 expected positives (42%) and 100 expected negatives (58%). The qualitative concordance of center 1–6 varied between 76% and 91% with a mean of 86% (Table 1, last columns). Centers 4 and 6 had the highest agreement with the consensus, whereas center 3 showed the lowest concordance (Table 1). Sensitivity for individual centers varied between 81% and 92%, whereas between centers there was considerable more variation in the specificity between centers ranging from 71 to 98% with an average of 86% (Table 1). Amongst all 388 positive data points, there were 56 false-negative assessments (14%). Among all the 520 single negatives, there were 74 false-positive assessments (14%). Fig. 1 shows the quantitative concordance of all LAIP results as observed by all six centers versus the expected values. Per center there are some differences in the agreement of their observed values with respect to the expected values, as shown by the discordant results, which were negative either by observation or by expectation that are depicted on the axes together with the indicated numbers (Fig. 1).

3.2. Impact of type of LAIP on qualitative concordance between centers

The kind of aberrancy present in the provided consensus LAIPs identified at diagnosis [17] is related to the outcome of concordance between centers (Table 2). Cross-lineage expression of LAIPs resulted in the highest qualitative concordance for almost all centers with a mean of 89%, followed by LAIPs with an asynchronous antigen expression (mean 85%, last columns Table 2). LAIPs of more mature myeloid cells, lacking CD34 and CD117, resulted in the lowest concordance (mean 79%). Especially, the latter had the highest percentage of false-negatives from all 75 positive estimates: 13% compared to 6% and 5% for asynchronous LAIPs and cross-lineage LAIPs, respectively. This results in the lowest sensitivity (72%) compared to asynchronous (85%) and cross-lineage (90%) LAIPS, respectively.

3.3. Analysis of MRD - qualitative

The next step was the interpretation of LAIP + cells to determine whether they do represent true MRD. The exchanged LMD were classified as MRD-positive (31/77; 40%) and MRD-negative (46/77; 60%). The qualitative concordance of center 1–6 varied between 77% and 96% with a mean of 85%. Centers 1 and 2 showed the highest agreement with the consensus, whereas center 4 showed the lowest concordance (Table 3). Between centers there was a substantial variation in sensitivity, ranging from 61% to 100% (mean 82%), but less variation in the specificity, ranging from 78 to 93% (mean 87%). From all 179 positive samples, there were 32 false-negative estimates (18%). Among all the 240 single negative samples, there were 31 false-positive estimates (13%) (Table 3).

Total agreement by all centers on the MRD-status was seen in 52% of samples overall (MRD-pos 45%; MRD-neg 55%). Agreement by at least 75% of centers that contributed (3 of 4, 4 of 5, or 5 of 6, since in part of the cases only 4 or 5 of the 6 centers reported data) was found in 75% of the total submitted cases. In only 8 patients there was disagreement between two centers and the other three or four centers. Eleven patients remained with an equal vote for either results (3 MRD positive, 3 MRD negative) on either side.

3.4. Analysis of MRD - quantitative

To determine concordance for the quantitative analyses of percentage MRD, we calculated the median value of all the individual reported percentages of each case and used this as the expected value, and subsequently compared this with the observed value reported by each center. Fig. 2 presents the differences of the six centers regarding the agreement in their observed versus expected values. ICC analysis resulted in an estimated correlation between individual centers of 0.62, indicating moderate similarity between observations within a AML case [18]. The estimated ICC between centers averaged over the six centers is high, 0.91. As shown in Fig. 2 there is reasonable variation between centers at low levels of MRD around the 0.1% cut-off, that is in samples with values ranging between 0.05 and 0.2%, and as a consequence resulting in false-positivity or false-negativity. From all 417 assessments, there were in total 66 (16%) discordant (MRD-positive vs. MRD-negative) between expected and observed quantitative MRD assessments. Ten of these 66 discrepancies (15%) were due to only minor differences of +/− 0.02% from expected quantitative values. Higher

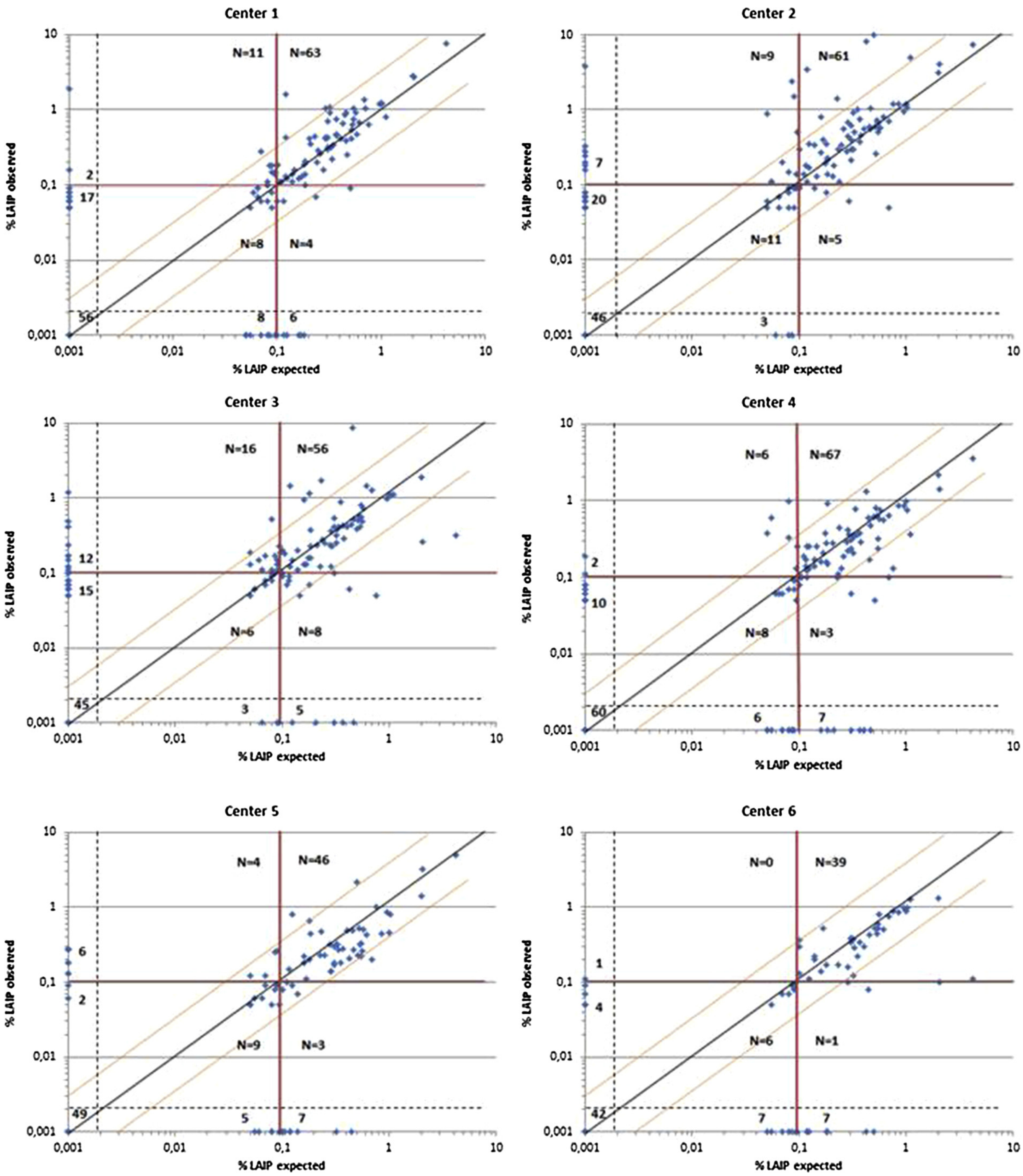

Fig. 1. Percentage LAIP-positive cells as observed by center 1 to 6 versus the expected values. The expected values were derived from the median of all estimates per sample. Each symbol represents one LAIP assessment. The diagonal lines in the quantitative part of the assay indicate the x = y, x = 3y and x = 0.33y axes. All LAIP percentages reported below 0.05% were considered as negative values and all concordantly negatives are cumulated on the intersection of x- and y-axis. Discordant results, which were negative either by observation or by expectation, are depicted on the axes. Numbers in the graphs represent the number of samples per indicated region.

levels of MRD above 0.2% are closer in distance to the identity line. As depicted in Fig. 2 almost all reported positive MRD percentages are within one-log variance. MRD-positive values outside one-log varied from 1 to a maximum of 3 outliers per center.

3.5. Impact of numbers of LAIPs on concordance in MRD percentage

Also the number of available LAIPs per AML follow up sample were of influence on the concordance between centers. In 17 AML cases only one LAIP and in 41 cases two LAIPs were identified at diagnosis, whereas in 19 cases at least three LAIPs were selected. As shown in Table 4 the overall agreement of 88% was the highest (center variation: 85%–93%) when two LAIPs were available for MRD interpretation. In those cases with only one LAIP available for interpretation, the agreement was only 77% (center variation: 72%–88%). The presence of at least three LAIPs did not result in a better agreement compared to one or two LAIPs.

Table 2
Concordance [a] in percentage LAIP + cells based on type of LAIPs.

Center	1		2		3		4		5		6		All		Range
exp-obs[b]	#	%	#	%	#	%	#	%	#	%	#	%	#	%	
Asynchronous antigen expression															
pos,pos	20	31	22	39	18	30	19	31	15	33	11	35	105	33	
neg,neg	37	58	27	48	24	40	37	61	21	46	18	58	164	52	
concordance		89		87		70		92		89		93		85	70-93
neg,pos[c]	3	7	6	18	13	35	0	0	8	28	0	0	30	15	
pos,neg[d]	4	17	1	4	5	22	5	21	2	12	2	15	19	15	
n =	64		56		60		61		46		31		318		
sensitivity[e]		83		96		78		79		88		85		85	78-96
specificity[f]		93		82		65		100		72		100		85	65-100
Crosslineage antigen expression															
pos,pos	35	49	29	41	28	39	31	44	20	37	19	42	162	42	
neg,neg	31	43	33	47	29	41	33	47	28	52	24	53	178	47	
concordance		92		88		80		91		89		95		89	80-95
neg,pos[c]	6	16	4	11	8	22	3	8	1	3	1	4	23	11	
pos,neg[d]	0	0	4	12	6	18	3	9	5	20	1	5	19	10	
n =	72		70		71		70		54		45		382		
sensitivity[e]		100		88		82		91		80		95		90	80-100
specificity[f]		84		89		78		92		97		96		89	78-97
Mature blasts															
pos,pos	7	26	12	46	11	42	10	37	8	32	6	23	54	34	
neg,neg	12	44	12	46	11	42	11	41	13	52	12	46	71	45	
concordance		70		92		84		78		84		69		79	69-92
neg,pos[c]	2	14	2	14	3	21	3	21	0	0	1	8	11	13	
pos,neg[d]	6	46	0	0	1	8	3	23	4	33	7	54	21	28	
n =	27		26		26		27		25		26		157		
sensitivity[e]		54		100		92		77		67		46		72	46-100
specificity[f]		86		86		79		79		100		92		87	79-100

[a] all percentages of LAIP equal or above cutoff (≥0.1%) are classified as 'positive', below (< 0.1%) as 'negative'.
[b] see method section for definition of expected versus observed result.
[c] false-positives = expected result is negative, whereas observation is positive.
[d] false-negatives = expected result is positive, whereas observation is negative.
[e] [(pos,pos)/[(pos,pos) + (pos,neg)]] * 100.
[f] [(neg,neg)/[(neg,pos) + (neg,neg)]] * 100.

4. Discussion

The present study was undertaken to evaluate the performance of laboratories in MRD assessment by MFC and especially to determine the concordance between those centers that participate in a multicenter setting. Moreover, the study was an attempt to establish critical issues that require further improvement for reaching a high degree of concordance in multicenter MRD assessment with MFC. Current MFC-based methodologies for detection of MRD depend on establishing a LAIP at diagnosis, and use this information at specified time points during or after therapy for detection of MRD [8,16,19], or to apply a standardized panel of antibody combinations for all MRD cases, irrespective the availability of a diagnosis sample, in a different-from-normal approach [20,21]. Our method made use of the first approach that provided a description of immunophenotypic abnormalities (LAIPs) for a given sample at diagnosis.

The observed qualitative concordance of the reported percentage of LAIP positive cells by the 6 participating centers was on average 86%. Performance of some centers were clearly better than others. Lowest concordance by center 3 (76%) was mainly due to the relative high number of observed false-positives resulting in a low specificity of 71%, whereas center 6 with the highest concordance reached a specificity of

Table 3
Qualitative [a] concordance in MRD assessment by the different centers.

Center	1		2		3		4		5		6		All		Range
exp,obs[b]	#	%	#	%	#	%	#	%	#	%	#	%	#	%	%
pos,pos	33	43	28	37	27	37	20	26	20	33	19	35	147	35	
neg,neg	41	53	40	53	31	42	39	51	31	52	27	50	209	50	
concordance		96		90		79		77		85		85		85	77-96
neg,pos[c]	3	7	4	9	9	22	5	11	5	14	4	13	31	13	
pos,neg[d]	0	0	4	12	6	18	13	39	4	17	4	17	32	18	
n =	77		76		73		77		60		54		417		
sensitivity[e]		100		88		82		61		83		83		82	61-100
specificity[f]		93		91		78		89		86		87		87	78-93

[a] all percentages of MRD equal or above cutoff (≥0.1%) are classified as 'positive', below (< 0.1%) as 'negative'.
[b] see method section for definition of expected versus observed result.
[c] false-positives = expected result is negative, whereas observation is positive.
[d] false-negatives = expected result is positive, whereas observation is negative.
[e] [(pos,pos)/[(pos,pos) + (pos,neg)]] * 100.
[f] [(neg,neg)/[(neg,pos) + (neg,neg)]] * 100.

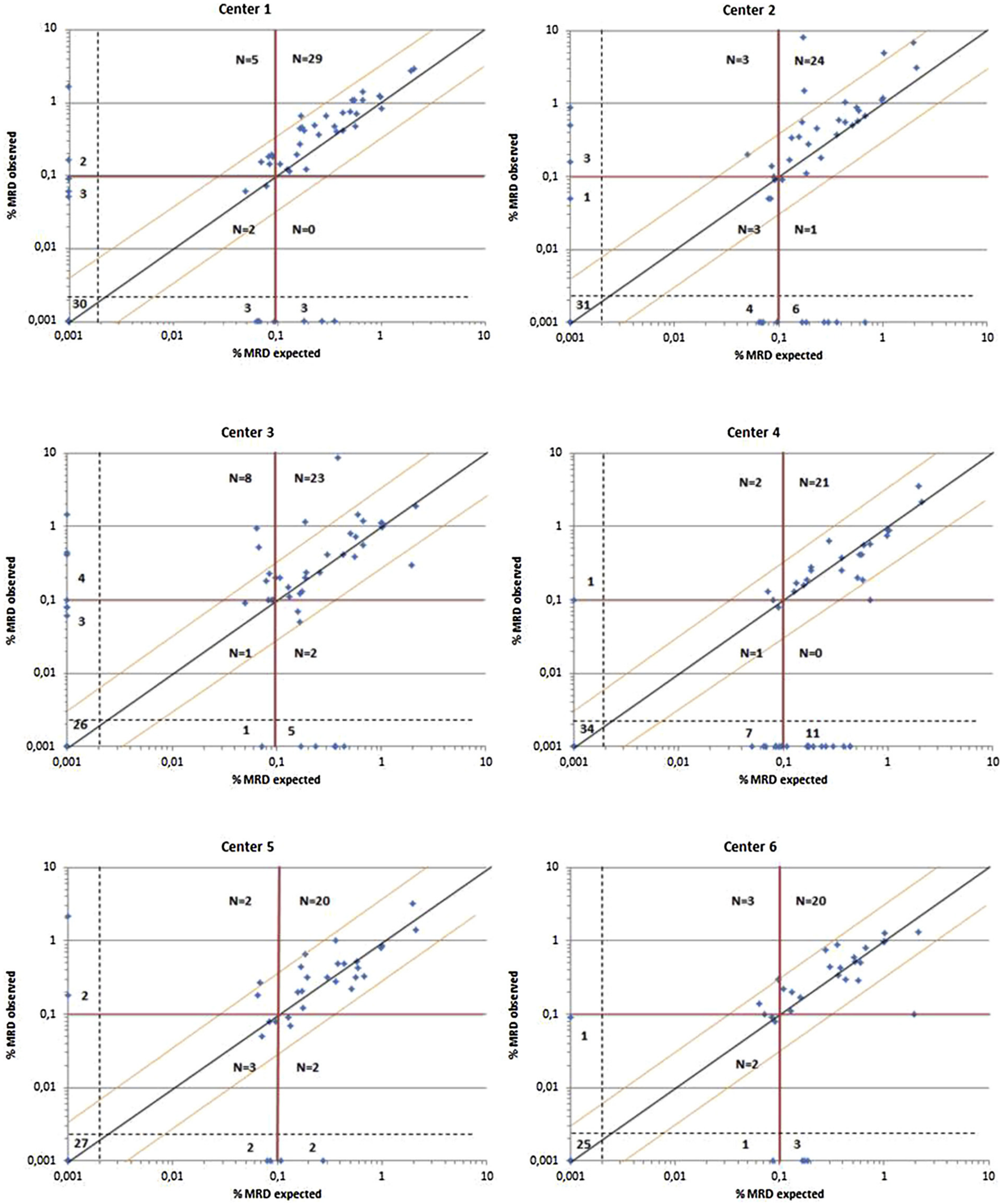

Fig. 2. Correlation analysis of MRD results as observed by center 1 to 6 versus the expected values. The expected values were derived from the median of all estimates per sample. Each symbol represents one MRD assessment. The diagonal lines in the quantitative part of the assay indicate the x = y, x = 3y and x = 0.33y axes. All MRD percentages reported below 0.05% were considered as negative values and all concordantly negatives are cumulated on the intersection of x- and y-axis. Discordant results, which were negative either by observation or by expectation, are depicted on the axes. Numbers in the graphs represent the number of samples per indicated region.

98%.

Also the type of LAIPs was of influence on the concordance between centers in detection of the percentage LAIP positive cells in follow-up samples. The highest concordance was observed for LAIP with cross-lineage expression (89%), followed by LAIPs with an asynchronous antigen expression. The lowest concordance (79%) was seen in those cases where no primitive marker expression was found (mature LAIP), resulting in a lower sensitivity as compared with the other type of LAIPs. Especially, in the latter the sensitivity varied considerable between centers, i.e. 46% to 100%, indicating that centers had difficulty with separating immunophenotypic aberrancy from regenerating BM, due to less experience with these type of mature LAIPs in combination with the fact that regenerating BM contains a higher percentage of immature cells. In particular, center 1, 5 and 6 reported a relative high

Table 4
Concordance[a] in MRD assessment based on number of available LAIPs.

Center	1		2		3		4		5		6		All		Range
exp-obs[b]	#	%	#	%	#	%	#	%	#	%	#	%	#	%	%
One LAIP															
pos,pos	7	41	5	31	6	43	6	35	4	27	4	29	32	34	
neg,neg	7	41	7	44	4	29	9	53	7	47	6	43	40	43	
concordance		82		75		72		88		74		72		77	72-88
neg,pos[c]	2	22	2	22	3	43	0	0	2	22	2	25	11	22	
pos,neg[d]	1	12	2	29	1	14	2	25	2	33	2	33	10	24	
n =	17		16		14		17		15		14		93		
sensitivity[e]		88		71		86		75		67		67		76	67-88
specificity[f]		78		78		57		100		78		75		78	57-100
Two LAIPs															
pos,pos	13	32	11	27	11	28	8	20	9	26	10	31	62	27	
neg,neg	25	61	26	63	23	58	25	61	22	65	19	59	140	61	
concordance		93		90		86		81		91		90		88	81-93
neg,pos[c]	3	11	2	7	4	15	3	11	2	8	2	9	16	10	
pos,neg[d]	0	0	2	15	2	15	5	38	1	10	1	9	11	15	
n =	41		41		40		41		34		32		229		
sensitivity[e]		100		85		85		62		90		91		85	62-100
specificity[f]		89		93		85		89		92		91		90	85-93
Three LAIPs															
pos,pos	9	47	9	47	7	37	6	32	6	55	4	50	41	43	
neg,neg	7	37	8	42	5	26	8	42	4	36	4	50	36	38	
concordance		74		89		63		74		91		100		81	63-100
neg,pos[c]	2	22	1	11	4	44	1	11	1	20	0	0	9	20	
pos,neg[d]	1	10	1	10	3	30	4	40	0	0	0	0	9	18	
n =	19		19		19		19		11		8		95		
sensitivity[e]		90		90		70		60		100		100		82	60-100
specificity[f]		78		89		56		89		80		100		80	56-100

[a] all percentages of LAIP cells equal or above cutoff (≥ 0.1%) are classified as 'positive', below (< 0.1%) as 'negative'.
[b] see method section for definition of expected versus observed result.
[c] false-positives = expected result is negative, whereas observation is positive.
[d] false-negatives = expected result is positive, whereas observation is negative.
[e] [(pos,pos)/[(pos,pos) + (pos,neg)]] * 100.
[f] [(neg,neg)/[(neg,pos) + (neg,neg)]] * 100.

number of false-negatives, resulting in a sensitivity of 54%, 67%, and 46%, respectively. Indeed, during the quarterly meetings of the working group, where we evaluated the performance of the participating institutes, it became clear that the mature type of LAIP resulted in more extensive discussions about MRD status than caused by the other type of LAIP. Overall, the quality of the LAIP is important for the reliability of the MRD assessment and depends not only on stability [22], but also on specificity. Specificity is determined by the degree of aberrancy in comparison with background expression in normal cells, and sensitivity depends on the percentage of blasts presenting with that LAIP at diagnosis [11,23,24]. The background expression of a LAIP can be determined on BM cells from healthy donors. However in clinical practice the expression of presumed leukemia-specific immunophenotypes can be very different in BM samples that are regenerating after chemotherapy [25]. In the multicenter approach of this study specificity for a particular LAIP was defined by each individual center, based on experience. Thus to classify LAIP events as MRD may differ between centers and therefore requires further standardisation.

Next, based on the observed percentage of LAIP positive cells the centers interpreted the available data whether or not these represent true MRD. The qualitative concordance in absence or presence of MRD of center 1–6 varied between 77% and 96% with a mean of 85%. The lowest concordance of 77% by center 4 is related to the low observed sensitivity by that center of 61%. This indicates that relatively high number of false-negatives were reported in comparison with the other centers. It shows that center 4 identified LAIP positive cells, but interpreted this as background in regenerating BM or not as a clear cluster of cells.

As expected, similar results were found when we compared the quantitative MRD concordance with the above qualitative MRD concordance. The blinded inter-laboratory tests of LMD interpretation showed overall a degree of quantitative concordance of 84% in MRD assessment (mean qualitative concordance 85%) and varied among the six centers between 75 and 89% (data not shown). The estimated ICC between individual centers indicates a moderate similarity within an AML case. A reason to find moderate similarity is disagreement of one or more of the six centers regarding an individual case. The causes of discordance in MRD assessment between centers, as discussed at our quarterly meetings, were mainly interpretation of background levels of LAIP in regenerating bone marrow, quality of LAIPs, and shifts in gate settings. Other causes in some cases were: MRD at limitation level of detection (not enough events), phenotypic shifts, and in one particular case in which autofluorescence of the analysed cells after treatment had effect on the gate settings. A second reason for a lower ICC is related to the variability in MRD values among subjects: the variability must be large to demonstrate reliability. A lack of such variability can occur when the MRD values are homogeneous (clustered to a small window of analysis) or in situations that one or more observers are stricter than others in their data analysis. In the present study the variability was not that large, since almost half of the follow-up samples that were assessed for MRD were reported as negative or zero (45% of total) and about a quarter were reported between 0.05% and 0.2% (27% of total). In cases with an observed percentage of LAIP positive cells around the cut-off of 0.1%, a lower agreement between centers is reached. In 15% of all expected versus observed discrepancies for quantitative MRD assessment were due to only minor variation of 0.02% in quantitation. This pinpoints to a serious limitation of a pre-defined cut-off: several studies have revealed that often a single cut-off level that defines MRD-positive and MRD-negative patient groups results on the one hand in a patient sub-group who classified as MRD-positive, but remained in complete remission. On the other hand, patient sub-groups who classified as MRD negative, but nevertheless relapsed, may be consistent with residual leukemia [15,16,26,27].

Our data also show that the number of available LAIPs per AML follow up sample were of influence on the concordance between centers. In those cases that only one LAIP was available for interpretation the agreement was only 77%. When however, two LAIPs were available for MRD interpretation the overall agreement was much higher (88%). The presence of at least three LAIPs resulted not in a better agreement than in the presence of only one or two LAIPs. One of the explanations could be that the third and more extra defined LAIPs are not based on additional aberrancies observed at diagnosis, but were in most cases overlapping variants of the first and second defined LAIP, so the same markers but with other fluorochrome combinations. Secondly, when more than two LAIPs are available it can occur that the percentages of LAIP positive cells are more often discordant among these LAIPs, and therefore making the interpretation more difficult. However, it is also important, whenever possible, to include more than one LAIP for MRD detection to prevent that post-therapeutic changes in immunophenotype are missed, e.g. due to selection by therapy of minor sub-populations. Such selections have been shown to occur for molecular clones [28], which may hypothetically be reflected in mutational and immunophenotypical changes [29,30].

To reach a high sensitivity it is essential to use highly informative antibody combinations in the LAIP. A disadvantages in this study was the use four-color antibody combinations that limits the possible antibody combinations and as a result limits the specificity and thereby the sensitivity of AML-specific events. Nowadays 8–10 colors are routinely used; this likely allows a more specific assessment of aberrancies. Nevertheless, the present study shows that a high variability is for a large part due to subjective interpretation, as is characteristic for the limitations of current MFC MRD protocols for AML.

An important advantage of using flow cytometric MRD assessment over PCR-based MRD detection in AML patients lies in the fact that MFC is applicable to virtually all patients [31–33], and is easily quantifiable with the additional ability to distinguish live from dead cells. Above that, it is considered as a less labor-intensive and faster MRD technique as compared to PCR. Nevertheless, it is logical that approaches combining molecular techniques and MFC also have been used, leading to improvement of MRD stratification in AML [34,35]. Next-generation sequencing (NGS) can be another application in the near future, since NGS enables the detection of molecular minimal residual disease in virtually every AML patient. A recent study where sequencing was compared with flow cytometry for the detection of residual disease showed that sequencing had significant additive prognostic value [36].

In conclusion, our study has shown that the analysis of acquired flowcytometric data to determine the level of MRD is a complex process that requires specific experience.

Nevertheless our results imply that immunophenotypic MRD assessment in AML will only be feasible when standardized methods are used for reliable multicentre assessment in large clinical trials. For that reason we think that more objective methods to identify and quantify aberrant cells will be necessary to finally allow implementation in standard diagnostic care.

Further improvement for reaching a higher degree of concordance in multicenter MRD assessment with MFC could be obtained by using: 1) multicolour immunostaining protocols using ≥ 8 colors that will likely go along with an improved specificity of MRD assessment by MFC; 2) standardization of every aspect of immunophenotypic MRD monitoring, including uniformity in the applied immunostaining protocols, antibody panels, acquisition [37,38], 3) automated gating strategies that probably have the largest impact on standardization. Several multidimensional analysis programs, originally developed for Cytof applications, are being adapted for MFC [39], including MRD [40–42]. Since MRD evaluation is becoming the new standard in evaluating response in AML [20,26,43], efforts need to be made by all laboratories involved in MRD detection to follow these unified standardized protocols, so that the results are comparable and reduce the inter-laboratory variability. In that respect, a consensus document that provides guidelines for MRD detection and its clinical use, has recently been published by the European Leukemia Net MRD working party [44].

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] B. Lowenberg, T. Pabst, J. Maertens, Y. van Norden, B.J. Biemond, H.C. Schouten, et al., Therapeutic value of clofarabine in younger and middle-aged (18-65 years) adults with newly diagnosed AML, Blood 129 (12) (2017) 1636–1645.

[2] C.S. Hourigan, M. Goswami, M. Battiwalla, A.J. Barrett, S. Sheela, J.E. Karp, et al., When the minimal becomes measurable, J. Clin. Oncol. 34 (21) (2016) 2557-+.

[3] B.D. Cheson, J.M. Bennett, K.J. Kopecky, T. Buchner, C.L. Willman, E.H. Estey, et al., Revised recommendations of the international working group for diagnosis, standardization of response criteria, treatment outcomes, and reporting standards for therapeutic trials in acute myeloid leukemia, J. Clin. Oncol. 21 (24) (2003) 4642–4649.

[4] A. Macedo, A. Orfao, M.B. Vidriales, M.C. Lopez-Berges, B. Valverde, M. Gonzalez, et al., Characterization of aberrant phenotypes in acute myeloblastic leukemia, Ann. Hematol. 70 (4) (1995) 189–194.

[5] M.C. Bene, J.S. Kaeda, How and why minimal residual disease studies are necessary in leukemia: a review from WP10 and WP12 of the European LeukaemiaNet, Haematologica 94 (8) (2009) 1135–1150.

[6] S. Schnittger, W. Kern, C. Tschulik, T. Weiss, F. Dicker, B. Falini, et al., Minimal residual disease levels assessed by NPM1 mutation-specific RQ-PCR provide important prognostic information in AML, Blood 114 (11) (2009) 2220–2231.

[7] N.M. Cruz, N. Mencia-Trinchant, D.C. Hassane, M.L. Guzman, Minimal residual disease in acute myelogenous leukemia, Int. J. Lab. Hematol. 39 (Suppl. 1) (2017) 53–60.

[8] D. Grimwade, P. Vyas, S. Freeman, Assessment of minimal residual disease in acute myeloid leukemia, Curr. Opin. Oncol. 22 (6) (2010) 656–663.

[9] J.F. San Miguel, M.B. Vidriales, C. Lopez-Berges, J. Diaz-Mediavilla, N. Gutierrez, C. Canizo, 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 98 (6) (2001) 1746–1751.

[10] A. Venditti, F. Buccisano, G. Del Poeta, L. Maurillo, A. Tamburini, C. Cox, et al., Level of minimal residual disease after consolidation therapy predicts outcome in acute myeloid leukemia, Blood 96 (12) (2000) 3948–3952.

[11] N. Feller, M.A. van der Pol, A. van Stijn, G.W. Weijers, A.H. Westra, B.W. Evertse, et al., MRD parameters using immunophenotypic detection methods are highly reliable in predicting survival in acute myeloid leukaemia, Leukemia 18 (8) (2004) 1380–1390.

[12] W. Kern, D. Voskova, C. Schoch, W. Hiddemann, S. Schnittger, T. Haferlach, 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 104 (10) (2004) 3078–3085.

[13] Group M-A-BS, C. Langebrake, U. Creutzig, M. Dworzak, O. Hrusak, E. Mejstrikova, et al., Residual disease monitoring in childhood acute myeloid leukemia by multiparameter flow cytometry: the MRD-AML-BFM Study Group, J. Clin. Oncol. 24 (22) (2006) 3686–3692.

[14] V.H. van der Velden, A. van der Sluijs-Geling, B.E. Gibson, J.G. te Marvelde, P.G. Hoogeveen, W.C. Hop, 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 24 (9) (2010) 1599–1606.

[15] S.D. Freeman, P. Virgo, S. Couzens, D. Grimwade, N. Russell, R.K. Hills, et al., Prognostic relevance of treatment response measured by flow cytometric residual disease detection in older patients with acute myeloid leukemia, J. Clin. Oncol. 31 (32) (2013) 4123–4131.

[16] M. Terwijn, W.L. van Putten, A. Kelder, V.H. van der Velden, R.A. Brooimans, T. Pabst, et al., High prognostic impact of flow cytometric minimal residual disease detection in acute myeloid leukemia: data from the HOVON/SAKK AML 42A study, J. Clin. Oncol. 31 (31) (2013) 3889–3897.

[17] N. Feller, V.H.J. van der Velden, R.A. Brooimans, N. Boeckx, F. Preijers, A. Kelder, et al., Defining consensus leukemia-associated immunophenotypes for detection of minimal residual disease in acute myeloid leukemia in a multicenter setting, Blood Cancer J. 3 (2013).

[18] P.E. Shrout, J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, Psychol. Bull. 86 (2) (1979) 420–428.

[19] G.J. Ossenkoppele, A.A. van de Loosdrecht, G.J. Schuurhuis, Review of the relevance of aberrant antigen expression by flow cytometry in myeloid neoplasms, Br. J. Haematol. 153 (4) (2011) 421–436.

[20] D. Grimwade, S.D. Freeman, Defining minimal residual disease in acute myeloid leukemia: which platforms are ready for "prime time"? Blood 124 (23) (2014) 3345–3355.

[21] E.L. Sievers, B.J. Lange, T.A. Alonzo, R.B. Gerbing, I.D. Bernstein, F.O. Smith, 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 101 (9) (2003) 3398–3406.

[22] M.R. Baer, C.C. Stewart, R.K. Dodge, G. Leget, N. Sule, K. Mrozek, 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 97 (11) (2001) 3574–3580.

[23] D. Voskova, S. Schnittger, C. Schoch, T. Haferlach, W. Kern, Use of five-color staining improves the sensitivity of multiparameter flow cytomeric assessment of minimal residual disease in patients with acute myeloid leukemia, Leuk. Lymphoma 48 (1) (2007) 80–88.

[24] B. Denys, A.J. van der Sluijs-Gelling, C. Homburg, C.E. van der Schoot, V. de Haas, J. Philippe, et al., Improved flow cytometric detection of minimal residual disease in childhood acute lymphoblastic leukemia, Leukemia 27 (3) (2013) 635–641.

[25] D. Olaru, L. Campos, P. Flandrin, N. Nadal, A. Duval, S. Chautard, et al., Multiparametric analysis of normal and postchemotherapy bone marrow: implication for the detection of leukemia-associated immunophenotypes, Cytometry B Clin. Cytom. 74 (1) (2008) 17–24.

[26] G. Ossenkoppele, G.J. Schuurhuis, MRD in AML: does it already guide therapy decision-making? Hematol. Am. Soc. Hematol. Educ. Program 2016 (1) (2016) 356–365.

[27] R.B. Walter, S.A. Buckley, J.M. Pagel, B.L. Wood, B.E. Storer, B.M. Sandmaier, et al., Significance of minimal residual disease before myeloablative allogeneic hematopoietic cell transplantation for AML in first and second complete remission, Blood 122 (10) (2013) 1813–1821.

[28] L. Ding, T.J. Ley, D.E. Larson, C.A. Miller, D.C. Koboldt, J.S. Welch, et al., Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing, Nature. 481 (7382) (2012) 506–510.

[29] C. Bachas, G.J. Schuurhuis, Y.G. Assaraf, Z.J. Kwidama, A. Kelder, F. Wouters, et al., The role of minor subpopulations within the leukemic blast compartment of AML patients at initial diagnosis in the development of relapse, Leukemia 26 (6) (2012) 1313–1320.

[30] W. Zeijlemaker, J.W. Gratama, G.J. Schuurhuis, Tumor heterogeneity makes AML a "moving target" for detection of residual disease, Cytometry B Clin. Cytom. 86 (1) (2014) 3–14.

[31] M.I. Del Principe, F. Buccisano, L. Maurillo, G. Sconocchia, M. Cefalo, M.I. Consalvo, et al., Minimal residual disease in acute myeloid leukemia of adults: determination, prognostic impact and clinical applications, Mediterr. J. Hematol. Infect. Dis. 8 (1) (2016) e2016052.

[32] D.M. Bahia, M. Yamamoto, L. Chauffaille Mde, E.Y. Kimura, J.O. Bordin, M.A. Filgueiras, et al., Aberrant phenotypes in acute myeloid leukemia: a high frequency and its clinical significance, Haematologica 86 (8) (2001) 801–806.

[33] W. Kern, S. Danhauser-Riedl, R. Ratei, S. Schnittger, C. Schoch, H.J. Kolb, et al., Detection of minimal residual disease in unselected patients with acute myeloid leukemia using multiparameter flow cytometry for definition of leukemia-associated immunophenotypes and determination of their frequencies in normal bone marrow, Haematologica 88 (6) (2003) 646–653.

[34] C. Marani, M. Clavio, R. Grasso, N. Colombo, F. Guolo, A. Kunkl, et al., Integrating post induction WT1 quantification and flow-cytometry results improves minimal residual disease stratification in acute myeloid leukemia, Leuk. Res. 37 (12) (2013) 1606–1611.

[35] G. Rossi, M.M. Minervini, L. Melillo, F. di Nardo, C. de Waure, P.R. Scalzulli, et al., Predictive role of minimal residual disease and log clearance in acute myeloid leukemia: a comparison between multiparameter flow cytometry and Wilm's tumor 1 levels, Ann. Hematol. 93 (7) (2014) 1149–1157.

[36] M. Jongen-Lavrencic, T. Grob, D. Hanekamp, F.G. Kavelaars, A. Al Hinai, A. Zeilemaker, et al., Molecular minimal residual disease in acute myeloid leukemia, N. Engl. J. Med. 378 (13) (2018) 1189–1199.

[37] T. Kalina, J. Flores-Montero, V.H.J. van der Velden, M. Martin-Ayuso, S. Bottcher, M. Ritgen, et al., EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols, Leukemia 26 (9) (2012) 1986–2010.

[38] P. Theunissen, E. Mejstrikova, L. Sedek, A.J. van der Sluijs-Gelling, G. Gaipa, M. Bartels, et al., Standardized flow cytometry for highly sensitive MRD measurements in B-cell acute lymphoblastic leukemia, Blood 129 (3) (2017) 347–357.

[39] L.M. Weber, M.D. Robinson, Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data, Cytometry A 89 (12) (2016) 1084–1096.

[40] S. Van Gassen, B. Callebaut, M.J. Van Helden, B.N. Lambrecht, P. Demeester, T. Dhaene, et al., FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data, Cytometry A 87 (7) (2015) 636–645.

[41] E. Coustan-Smith, G. Song, S. Shurtleff, A.E. Yeoh, W.J. Chng, S.P. Chen, et al., Universal monitoring of minimal residual disease in acute myeloid leukemia, JCI Insight 3 (9) (2018).

[42] J. Flores-Montero, L. Sanoja-Flores, B. Paiva, N. Puig, O. Garcia-Sanchez, S. Bottcher, et al., Next Generation Flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma, Leukemia 31 (10) (2017) 2094–2103.

[43] E. Paietta, Should minimal residual disease guide therapy in AML? Best Pract. Res. Clin. Haematol. 28 (2-3) (2015) 98–105.

[44] G.J. Schuurhuis, M. Heuser, S. Freeman, M.C. Bene, F. Buccisano, J. Cloos, et al., Minimal/measurable residual disease in AML: consensus document from ELN MRD Working Party, Blood 131 (12) (2018) 1275–1291.