Bone Marrow Transplantation (2020) 55:843–850
https://doi.org/10.1038/s41409-019-0729-4

EDITORIAL

Statistics and measurable residual disease (MRD) testing: uses and abuses in hematopoietic cell transplantation

Megan Othus[1] · Robert Peter Gale[2] · Christopher S. Hourigan[3] · Roland B. Walter[4,5,6,7]

Received: 6 October 2019 / Revised: 9 October 2019 / Accepted: 15 October 2019 / Published online: 30 October 2019

Series Editors' Note

The decision whether to recommend a transplant to someone with acute leukemia in first remission is complex and challenging. Diverse, often confounded co-variates interact to influence one's recommendation. Briefly, the decision metric can be viewed in three spheres: (1) subject-; (2) transplant-; and (3) disease-related co-variates. Subject-related co-variates include items such as age and comorbidities. Transplant-related co-variates include items such as donor-types, graft source, proposed conditioning and pre- and post-transplant immune suppression.

But what of disease-related variables? Previously haematologists relied on co-variates such as WBC at diagnosis, chemotherapy cycles to achieve first remission, cytogenetics and most recently, mutation topography. However, these co-variates have largely been replaced by results of measurable residual disease (MRD)-testing. Many chemotherapy-only and transplant studies report strong correlations between results of MRD-testing on therapy outcomes such as cumulative incidence of relapse (CIR), leukemia-free survival (LFS) or survival. (CIR makes biological sense in a transplant context whereas LFS and survival do not give competing causes of death such as transplant-related mortality (TRM), graft-versus-host disease and interstitial pneumonia unrelated to relapse probability).

This raises the question of how useful results are of MRD-testing in predicting CIR after transplants. Elsewhere we discussed accuracy and precision of MRD-testing in predicting outcomes of therapy of acute myeloid leukemia (Estey E, Gale RP. Leukemia 31:1255−1258, 2017; Hourigan CS, Gale RP, Gormley NJ, Ossenkoppele GJ, Walter RB. Leukemia 31:1482−1490, 2017). Briefly put, not terribly good. Although results of MRD-testing are often the most powerful predictor of CIR in multivariable analyses, the C-statistic (a measure of prediction accuracy) is often only about 0.75. This is much better than flipping a *fair coin* but far from ideal.

In the typescript which follows, Othus and colleagues discuss statistical issues underlying MRD-testing in the context of haematopoietic cell transplants. We hope readers, especially haematologists who often need to make transplant recommendations to people with acute leukemia in first remission, will read it carefully and critically. The bottom line is MRD-test data are useful but considerable uncertainty is unavoidable with substantial false-positive and -negative rates. We need to acknowledge this uncertainty to ourselves and to the people we counsel. The authors quote Voltaire who said: *Doubt is not a pleasant condition, but certainty is an absurd one*. Sadly so, but we do the best we can.

Robert Peter Gale, Imperial College London, and Mei-Jie Zhang, Medical College of Wisconsin and CIBMTR.

✉ Roland B. Walter
rwalter@fredhutch.org

[1] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[2] Haematology Research Centre, Division of Experimental Medicine, Department of Medicine, Imperial College London, London, UK

[3] Myeloid Malignancies Section, Hematology Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA

[4] Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[5] Department of Medicine, Division of Hematology, University of Washington, Seattle, WA, USA

[6] Department of Pathology, University of Washington, Seattle, WA, USA

[7] Department of Epidemiology, University of Washington, Seattle, WA, USA

> Doubt is not a pleasant condition, but certainty is an absurd one.
>
> Voltaire, in: *Letter to Frederick William, Prince of Prussia*; 28 November 1770

Introduction

Testing for measurable ('minimal') residual disease (MRD) in people with acute leukaemias and other haematologic cancers has gained popularity [1–16]. Results of these tests are now often included as an endpoint in reports of clinical trials outcomes and increasingly used in clinical practice with haematopoietic cell transplantation no exception. In addition to stratification for the risk of cancer recurrence, MRD testing is used to inform transplant-related medical decisions. For example, many experts, consensus statements, and management guidelines suggest considering results of MRD testing in the decision whether persons with acute lymphoblastic leukaemia (ALL) or acute myeloid leukaemia (AML) should receive a transplant in first remission, in selecting the type of haematopoietic cell graft, intensity of pretransplant conditioning and type, intensity or duration of post-transplant interventions such as immune suppression and/or pre-emptive post-transplant anti-leukaemia therapy [17–21]. However, as with any other prognostic or predictive test, the interpretation of MRD-test results is subject to limitations in statistical properties that need to be considered when translating these data to the clinic. Adding to this complexity, there are many techniques to quantify MRD and each has different operating characteristics. Treatment strategies are often decided based on results of one MRD test reported as a binary (negative or positive). This approach ignores basic characteristics of these tests, and the test's accuracy and precision (Fig. 1) in predicting clinical outcomes is not well described and often misunderstood. MRD tests using different techniques, such as multi-parameter flow cytometry and quantitative polymerase chain reaction (PCR), done on the same sample may give different results, especially when the readout is a binary: positive or negative [22, 23]. As such, data from MRD tests using different techniques should be considered complementary rather than duplicative. Discordances further complicate interpreting MRD-test results. Using binary readouts from MRD tests has several statistical issues besides decreased sensitivity and specificity including decreased power, underestimation of variation in outcome between groups (persons with very low-level positive MRD-test levels may be outcome-wise closer to MRD-test-negative persons than those who test high-level MRD positive), and inability to identify any linear relationships with outcomes [24]. Flexible models of quantitative MRD, such as spline models ([25]; elaborated on in another part of this series), can help to elucidate non-linear relationships in the data.

Here we discuss characteristics and properties common to all MRD tests including sensitivity, specificity, accuracy, precision, and positive- and negative-predictive values (see Table 1 for glossary and definitions of statistical terms). We define and compare these quantities, discuss their role in informing medical decisions, and describe the role of randomized trials in evaluating MRD-test results. For a broader discussion of outcome prediction in people with haematologic cancers, see [26].

The perfect MRD test and why it does not yet exist (and may never exist)

Critical appraisal of the performance properties of any MRD test requires defining what the ideal MRD test should do and what is the clinical comparator. For example, is the test designed to detect some or all residual cancer cell(s) or only cancer cells *biologically* able to cause relapse within a specified interval (perhaps the subject's remaining lifetime) or which *cause* relapse within a specified observation interval? These are distinct, sometimes overlapping, goals. Equally relevant is the outcome we want to predict with the MRD-test result. Do we plan to use it to predict relapse, best analysed as cumulative incidence of relapse (CIR) because of competing events, relapse-free or event-free survival, overall survival, or some other endpoint of clinical interest? With the current interest in MRD testing for risk-stratification and treatment decision-making, a perfect MRD

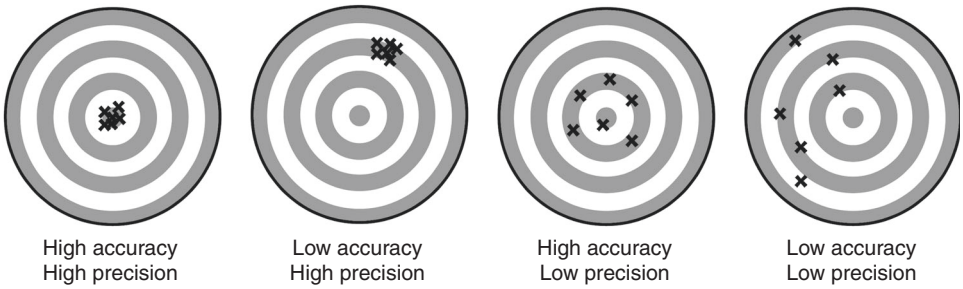

Fig. 1 Accuracy (closeness of repeat measurements to true value) and precision (closeness of repeat measurements to each other) of medical tests (e.g. MRD tests)

Table 1 Glossary of statistical terms

Accuracy	Closeness of a measurements to true value
AUC—area under the receiver operating characteristic (ROC) curve	AUC is equal to the probability a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative')
C-statistic	Probability someone with an event had a higher risk score than a someone without an event
Negative-predictive value	Proportion of negative test results that are true-negatives
Power	Probability a test rejects the null hypothesis when a specific alternative hypothesis is true
Precision	Closeness of repeat measurements to each other
Positive predictive value	Proportion of positive test results which are true-positives
Receiver operating characteristic (ROC) curve	Graphical plot displaying the sensitivity and one-specificity ability of a quantitative score to classify a binary outcome as the score's discrimination threshold is varied
Repeatability (test–retest reliability)	Closeness of the agreement between the results of successive measurements of the same measure carried out under the same condition (same person, same experimental setup)
Replicability	Closeness of the agreement between the results of measurements of the same measure carried with same method (different person, same experimental setup)
Reproducibility	Closeness of the agreement between the results of measurements of the same measure carried with different method (different person, different experimental setup)
Sensitivity	True-positive rate—proportion of true-positives that are found to be MRD positive
Specificity	True-negative rate—proportion of true-negatives that are found to be MRD negative

test might be defined as one which accurately identifies and quantifies the smallest population(s) of cancer cells in someone in histological complete remission which, if left untreated, cause relapse whilst being indifferent towards residual cancer cells which do not cause relapse during a specified observation interval.

The follow-up period in retrospective MRD-test analyses is important but often ignored. Most clinical trials have a finite follow-up interval, say, 2 or 5 years. We can evaluate whether results of a positive MRD test predict relapse over a lifetime only if all relapses are observed during this interval (i.e. there will be no further relapses after the finite follow-up time). Are the positive MRD tests in subjects without relapse in the follow-up interval false-positives or has the observation interval been insufficient for all relapses to occur? Moreover, it is likely some subjects who might have relapsed during the observation interval died of other causes, say graft-versus-host disease (GvHD) or a heart attack, before they could relapse. We can correct for this only imperfectly by accounting for competing causes of failure as is done by CIR analyses. Other persons may die after the observation interval of related or unrelated events, say cancer recurrence or a stroke. We will not, of course, know of these events. As such, there must be an unavoidable rate of false-positive and -negative MRD-test results, real or not, when events occur beyond the observation interval if the goal is to use MRD to evaluate lifetime risk. Further complicating this analysis is that some patients may have such high risk of a competing event (e.g. non-relapse mortality) that their CIR will not be relevant as such patient would be predicted to die before relapse could ever occur.

Although several technologies focused on immune phenotype or cytogenetic and/or molecular abnormalities have been developed to detect neoplastic haematopoietic cells, each with advantages and disadvantages, our understanding of cancer stem cells and how they differ in the context of immune phenotype and/or molecular features from other cancer cell populations is incomplete and unavoidably imperfect [3–6, 9–11, 27–33]. In other words, one reason the perfect MRD test with 100% sensitivity and specificity to predict cancer recurrence at the cohort- or subject-levels does not yet exist is related to incomplete knowledge of the neoplastic cells able to cause relapse. There are however additional reasons accounting for the substantial rates of false-positive and -negative tests.

For the clinical performance of any MRD test the theoretical maximum sensitivity and specificity of an assay to detect operationally relevant residual cancer cells (i.e. those causing relapse), together with other characteristics such as the reproducibility and repeatability or test–retest reliability (the components of a test's precision) or replicability are important. Of course, these characteristics are not unique to MRD tests but apply to other medical assessments such as the histological assessment of a bone marrow specimen [34]. The precision of the test may be impacted by small volume sample (discussed later) as well as measuring technology (e.g. calibration of flow cytometers or PCR machines). In addition to sampling site and volume, other sampling details (timing, frequency, etc.) and result interpretation, for which many uncertainties remain, are of practical consideration. For example, even using histological criteria for complete remission, synchronous biopsies at

several sites may be discordant. This is true not only for solid cancers such as prostate cancer but also for haematopoietic cancers. Discordance rates are even higher for metachronous biopsies. In other words, the perfect MRD test to predict cancer recurrence at the cohort- or subject-levels may never exist.

Measures of accuracy for binary definitions of MRD

When results of a test with quantitative, numerical data are reduced to a binary outcome, an unfortunate and often inaccurate strategy, there are four possible, mutually exclusive, outcomes: (1) true-positive; (2) true-negative; (3) false-positive; or (4) false-negative. The 2×2 diagram in Table 2 summarizes the distribution of the four test outcomes. In medical testing we are often concerned with falsepositives and negatives. Assume in this example we wanted to develop an MRD test that would identify and quantify cells which without further treatment are biologically able to cause relapse within a specified interval. For such a test, a false-positive would be an MRD test result indicating there are remaining cancer cells destined to cause relapse when, in fact, no intervention is needed to prevent cancer recurrence within the specified observation interval. This could be because the test is not sufficiently specific or because it identifies cancer cells that cannot or do not cause relapse within the observation interval. There are several potential reasons for these errors including the cells detected by the MRD test lack the *biological ability* to cause cancer recurrence during the observation interval or because of stochastic considerations (the cells have the biological ability to cause relapse but this does not occur for unpredictable reasons such as the cell(s) never divide(s)). A false-negative MRD test would indicate that there were no remaining cells which would result in relapse unless there is an effective intervention. The true-positive rate is equal to $1 -$ the false-negative rate and is often referred to as sensitivity. The true-negative rate is equal to $1 -$ the false-positive rate and is often referred to as specificity.

In addition to the four measures described above, positive- and negative-predictive values (PPV and NPV) are important in understanding the performance of any test. PPV and NPV are the proportions of positive tests which are true-positives and negative tests which are true-negatives. PPV and NPV depend on sensitivity and specificity of the test and, importantly, on the *true* prevalence of positive subjects compared with the positive and negative test results. These values can be estimated for a binary test and binary outcome with straightforward 2×2 table calculations. Table 3 provides sample data showing how these values can be calculated. In this example we use MRD measured by multi-parameter flow cytometry from an AML trial in subjects age 18–60 years [35, 36]. We note similar calculations can be done with more complicated definitions of MRD including combining MRD results across multiple time-points or summaries of MRD kinetics over time.

In this cohort MRD data were available on 170 subjects achieving histological complete remission [36]. Relapse-free survival at 1 year was measured from the date of histological complete remission and relapse and death were considered events. In this example:

- sensitivity (true-positive rate = probability MRD test is positive amongst subjects who will relapse and/or die in the following year) $= \frac{18}{18+20} = 47\%$
- specificity (true-negative rate = probability MRD test is negative amongst subjects who will neither relapse nor die in the following year) $= \frac{99}{99+33} = 75\%$
- PPV (probability of experiencing relapse and/or death within 1 year amongst subjects who are MRD positive) $= \frac{18}{33+18} = 35\%$
- NPV (probability of experiencing neither relapse nor death within 1 year amongst subjects who are MRD negative) $= \frac{99}{99+20} = 83\%$

Even acknowledging additional anti-leukaemia therapy was given before the 1-year mark to most of these subjects, these calculations highlight data from this MRD test with the outcome of relapse at 1-year result in substantial mis-classification rates. However, even with this high level of mis-classification, the MRD-test result is strongly associated with relapse-free survival with an odds ratio of 2.7 for 1-year relapse-free survival consistent with the strong prognostic association of MRD observed across many cohorts of people with AML [5].

Table 2 Classification of MRD-test results by true relapse state without further treatment

	True relapse state	
	Will not relapse	Will relapse
MRD-test negative	True-negative	False-negative
MRD-test positive	False-positive	True-positive

Table 3 MRD-test results by relapse state

	True relapse state	
	No RFS event in 1 year	RFS event in 1 year
MRD-test negative	99	20
MRD-test positive	33	18

MRD was retrospectively evaluated [35]. Relapse-free survival (RFS) was measured from the date of histological complete remission to the first of either relapse or death

Unfortunately, sensitivity, specificity, PPV, and NPV of MRD tests are not routinely described in biomedical publications. Most focus on the prognostic strength of the MRD test showing, on average, outcomes of persons with MRD-negative tests are significantly better than outcomes of persons with MRD-positive tests. Typically, the outcome interrogated is survival although an MRD-test result is biologically more likely to correlate with CIR because survival is influenced by other outcomes, including some, such as therapy-related toxicity (TRM) not expected to correlate with MRD-test result, whereas others like GvHD are confounded with CIR (persons with G*v*HD are less likely to relapse than those without G*v*HD). Although understandable from a clinical perspective (and perhaps driven by requirements from Health Authorities) the focus of many if not most reports on survival rather than CIR makes little biological sense. As an additional limitation, when estimating sensitivity, specificity, or other statistical quantities in settings with censored data or competing risks, a 2×2 table cannot be accurately constructed and specific statistical methodologies are needed to account for these data features [37]. These analyses are complex and debatable; for example, the definition of specificity can vary on how persons with a competing event are analysed [38].

Because of the strong correlation between MRD-test results and cancer recurrence (and related outcomes) and the diverse treatment options for many persons with haematologic cancers many physicians wish to use MRD-test results to determine best-possible therapy options. But biomarkers with strong prognostic associations can still have very poor predictive properties with respect to identifying the *best* therapy for someone [39]. For example, an odds ratio (OR) of 3 can be associated with greater than 50% false-positive or false-negative rates. A test with 90% specificity and 80% sensitivity can require an OR of 36 or higher, much higher than odds ratios or hazard ratios typically found in reports of MRD testing in the biomedical literature. But even an OR of 36 may be insufficient to have a very accurate biomarker. For example, a sensitivity of 97% and a specificity of 50% will also have an odds ratio of 36, highlighting the importance of reporting values including sensitivity and specificity, not just an OR or a hazard ratio.

Accuracy of quantifying MRD-test data

Although MRD tests are typically quantitative, results are often reported as positive or negative. This quantitative measurement can be converted to a binary measurement (positive or negative) by identifying people as positive who have any residual cancer cells detected by the test or by setting a minimum threshold of residual disease to be detected, for example, >0.1% of cells with an abnormal immune phenotype or residual cells with a mutation variant allele frequency >0.001. Many statistical methods are proposed to identify the 'best' threshold for creating a binary but using thresholds instead of the quantitative measurement is often associated with reduced (at times, substantially reduced) predictive performance [25, 40].

A common statistic reported as a generalization of sensitivity or specificity is the area under the receiver operating characteristic (ROC) curve (AUC) often translated to a concordance or C-statistic, discussed below. The ROC curve is plotted by tabulating the sensitivity and one-specificity of binary markers defined by every possible cut-point in the quantitative biomarker. As such it is invariant to the scale or units of the biomarker and so ROCs can be compared for different MRD measurements. The AUC is a single-summary value of the ROC and is constrained to be between 0 and 1. The C-statistic is the proportion of pairs of persons correctly ranked by the biomarker (i.e. the person with worse outcome in the pair also has a worse biomarker score). A C-statistic <0.5 is consistent with the test being worse than the flip of a *fair coin* in predicting outcome, a C-statistic of 0.5 is consistent with a flip of a *fair coin* and a C-statistic of 1 is perfect prediction (no false-positives or negatives). With a binary outcome, the AUC is equal to the C-statistic. C-statistics are defined more generally than AUC and so can be reported for time-to-event endpoints. The C-statistic for the data in Table 2, 0.59, shows weak predictive accuracy. We note that the C-statistic is a function of the prevalence of a biomarker, and so a test with constant sensitivity and specificity can vary between populations with varying MRD-positive prevalence.

What even the perfect MRD test cannot tell you

It is unlikely any MRD test will have perfect sensitivity to detect a designated target or targets and/or biomarker(s). Even with perfect sensitivity one might get a false-negative MRD-test result because of inconsistent presence of the assay target(s) in cancer cells.

As we summarized previously [5], it is a common misunderstanding that improvements in the MRD-test technology will eventually eliminate false-negative MRD tests by providing a complete accounting of the remaining residual cancer cells. Rather, the ability to detect low levels of residual cancer cells is limited primarily by the character and size of the sample tested, not MRD-test sensitivity. This is an important limitation considering MRD tests are typically based on small samples such as 1 mL of bone marrow from an estimated 750 mL volume in a 70 kg male or a 10 mL blood sample from a 5.5 L estimated blood volume. There is also the issue of topographic heterogeneity. For

example, leukaemia cells are thought to occupy specific bone marrow *niches* rather than being uniformly distributed. Taking larger bone marrow samples will not necessarily resolve this bias. For example, bone marrow samples larger than 5 mL simply contain more blood cells, not more bone marrow cells [41, 42].

The MRD-test performance may also be impacted by the frequency of testing. Single time-point measurements more likely result in false-positive and -negative MRD-test results than when result trends are considered. Re-testing can decrease the likelihood of incorrectly interpreting an MRD-test result. Even without intervention, however, repeat MRD-testing results are occasionally discordant in both possible directions: a negative to positive MRD test or the converse. Discordances have many explanations but precision of the test and small volume sampling of a topographically heterogeneous population of cancer cells [43, 44] are important considerations. Requiring concordant results to declare a person MRD-positive or -negative increases specificity but decreases sensitivity. Test result validation using alternative methodologies may be useful in such circumstance but may be impractical. Sequential MRD testing can be particularly helpful as a strategy to increase sensitivity if changes in MRD-levels (e.g. increasing transcript levels or increasing percentage of immune-phenotypically abnormal cells) are the readout. A single discordant datapoint would be insufficient to make an estimate of clinically relevant changes in residual cancer cells. The optimal interval and duration of sequential MRD testing is unknown and may depend on several variables such as the type and mutation profile of the cancer, type of therapy, or interval since achieving remission.

Evidence for basing treatment decisions on MRD-test results

Biomarkers measured prior to treatment to provide information of the likelihood of response to a specific therapy are often called *predictive* biomarkers. Common examples of predictive biomarkers in oncology are genomic biomarkers which indicate who should receive *a targeted therapy*, or 'fitness' biomarkers such as a performance score indicating a persons' ability to survive intensive therapy. Because MRD tests are imperfect, because people are misclassified and because retrospective and observational studies are subject to many biases [45], randomized trials are needed to prove the benefit of using MRD-test results for therapy decisions such as whether or not to do a transplant in someone with ALL or AML in first remission. Not only because MRD-test technologies are quickly evolving, there has been insufficient commitment to the large, long-term randomized clinical trials needed to prove the value in making therapy decisions based on results of MRD testing. Nevertheless, such trials are needed to accurately characterize the trade-offs present in such a treatment strategy. Retrospective non-randomized data cannot be used to evaluate this because the outcomes are often confounded by physician actions based on the results of MRD tests such as giving additional therapy or performing a transplant or making treatment decisions on the basis on other criteria including clinical judgement.

False-negatives and -positives are important when considering MRD-test results for decisions on interventions associated with serious adverse medical consequences such as allogeneic hematopoietic cell transplant. Assuming an example in which a cohort of AML patients underwent MRD testing at the completion of post-remission therapy (Fig. 2; hypothetical example like data reported by Terwijn et al. [46]):

- If all MRD-test-positive persons were deemed at high risk of relapse and received transplant, many persons would have received it with no possibility of benefit and substantial possibility of harm.
- If all MRD-test-negative persons were deemed low risk of relapse and did not receive a transplant, many would have relapsed. Whether or not they could be *rescued* with a transplant at this time is controversial. However, there are no convincing data a transplant done earlier would have improved their subsequent relapse outcome.

Our bottom line

Interpreting results of MRD testing is complex. Whether making therapy decisions based on MRD-test results

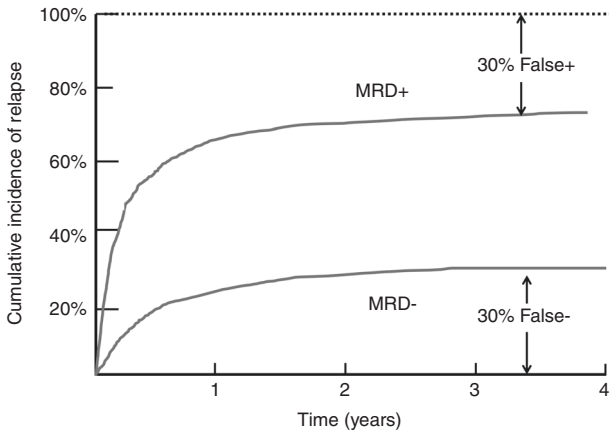

Fig. 2 Hypothetical example of cumulative incidence of relapse for adults with AML with positive or negative MRD-test result after completion of post-remission therapy. Abstracted from data reported by Terwijn et al. [46]

improves clinical outcomes can only be tested in randomized clinical trials [47–50]. Such data are lacking. Is MRD testing useful? Clearly. However, it is important physicians understand when they act on an MRD-test result, either by giving or withholding a therapy, they will often be wrong. Risks associated with these MRD-test result-based decisions are asymmetrical. When the intervention is safe, an incorrect prediction may have little medical consequence although it may have other adverse effects such as psychological and fiscal. In contrast, when the intervention is associated with serious adverse medical consequences including death, this uncertainty needs to be acknowledged by the physician and conveyed to the patient. Harm may be of a lesser magnitude when the decision to withhold an intervention is based on results of MRD testing as there are no convincing data yet that most earlier interventions for an event such as cancer recurrence improves outcomes [51–53]. As Stephen Hawking said: *The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge*.

Acknowledgements RPG acknowledges support from the National Institute of Health Research (NIHR) Biomedical Research Centre funding scheme. This work was supported in part by the intramural programme of the National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health (NIH) and by the following PHS/DHHS grant numbers awarded by the National Cancer Institute (NCI), National Clinical Trials Network (NCTN) to SWOG: CA180888 and CA180819.

Compliance with ethical standards

Conflict of interest CSH receives research funding from Merck Sharpe & Dohme and SELLAS Life Sciences Group AG. RPG is a part-time employee of Celgene Corp. Opinions expressed herein are those of the authors and do not represent the official position of the National Institutes of Health, US Food and Drug Administration or the United States government. The other authors declare that they have no conflict of interest.

References

1. Campana D, Pui CH. Minimal residual disease-guided therapy in childhood acute lymphoblastic leukemia. Blood. 2017;129:1913–8.
2. Bassan R, Intermesoli T, Scattolin A, Viero P, Maino E, Sancetta R, et al. Minimal residual disease assessment and risk-based therapy in acute lymphoblastic leukemia. Clin Lymphoma Myeloma Leuk. 2017;17S:S2–9.
3. Brüggemann M, Kotrova M. Minimal residual disease in adult ALL: technical aspects and implications for correct clinical interpretation. Blood Adv. 2017;1:2456–66.
4. Della Starza I, Chiaretti S, De Propris MS, Elia L, Cavalli M, De Novi LA, et al. Minimal residual disease in acute lymphoblastic leukemia: technical and clinical advances. Front Oncol. 2019;9:726.
5. Hourigan CS, Gale RP, Gormley NJ, Ossenkoppele GJ, Walter RB. Measurable residual disease testing in acute myeloid leukaemia. Leukemia. 2017;31:1482–90.
6. Schuurhuis GJ, Heuser M, Freeman S, Béné MC, Buccisano F, Cloos J, et al. Minimal/measurable residual disease in AML: a consensus document from the European LeukemiaNet MRD Working Party. Blood. 2018;131:1275–91.
7. Thompson PA, Wierda WG. Eliminating minimal residual disease as a therapeutic end point: working toward cure for patients with CLL. Blood. 2016;127:279–86.
8. Owen C, Christofides A, Johnson N, Lawrence T, MacDonald D, Ward C. Use of minimal residual disease assessment in the treatment of chronic lymphocytic leukemia. Leuk Lymphoma. 2017;58:2777–85.
9. Tomuleasa C, Selicean C, Cismas S, Jurj A, Marian M, Dima D, et al. Minimal residual disease in chronic lymphocytic leukemia: a consensus paper that presents the clinical impact of the presently available laboratory approaches. Crit Rev Clin Lab Sci. 2018;55:329–45.
10. Holstein SA, Avet-Loiseau H, Hahn T, Ho CM, Lohr JG, Munshi NC, et al. BMT CTN Myeloma Intergroup workshop on minimal residual disease and immune profiling: summary and recommendations from the organizing committee. Biol Blood Marrow Transpl. 2018;24:641–8.
11. Soverini S, De Benedittis C, Mancini M, Martinelli G. Best practices in chronic myeloid leukemia monitoring and management. Oncologist. 2016;21:626–33.
12. Tantravahi SK, Guthula RS, O'Hare T, Deininger MW. Minimal residual disease eradication in CML: does it really matter? Curr Hematol Malig Rep. 2017;12:495–505.
13. Landgren O, Lu SX, Hultcrantz M. MRD testing in multiple myeloma: the main future driver for modern tailored treatment. Semin Hematol. 2018;55:44–50.
14. Berger N, Kim-Schulze S, Parekh S. Minimal residual disease in multiple myeloma: impact on response assessment, prognosis and tumor heterogeneity. Adv Exp Med Biol. 2018;1100:141–59.
15. Kothari S, Hillengass J, McCarthy PL, Holstein SA. Determination of minimal residual disease in multiple myeloma: does it matter? Curr Hematol Malig Rep. 2019;14:39–46.
16. Bal S, Weaver A, Cornell RF, Costa LJ. Challenges and opportunities in the assessment of measurable residual disease in multiple myeloma. Br J Haematol. 2019;186:807–19.
17. Athale UH, Gibson PJ, Bradley NM, Malkin DM, Hitzler J, Group PMW. Minimal residual disease and childhood leukemia: standard of care recommendations from the Pediatric Oncology Group of Ontario MRD Working Group. Pediatr Blood Cancer. 2016;63:973–82.
18. DeFilipp Z, Advani AS, Bachanova V, Cassaday RD, Deangelo DJ, Kebriaei P, et al. Hematopoietic cell transplantation in the treatment of adult acute lymphoblastic leukemia: updated 2019 evidence-based review from the American Society for Transplantation and Cellular Therapy. Biol Blood Marrow Transplant. 2019 pii: S1083-8791(19)30528-2. https://doi.org/10.1016/j.bbmt.2019.08.014 [Epub ahead of print].
19. Giebel S, Marks DI, Boissel N, Baron F, Chiaretti S, Ciceri F, et al. Hematopoietic stem cell transplantation for adults with Philadelphia chromosome-negative acute lymphoblastic leukemia in first remission: a position statement of the European Working Group for Adult Acute Lymphoblastic Leukemia (EWALL) and the Acute Leukemia Working Party of the European Society for Blood and Marrow Transplantation (EBMT). Bone Marrow Transpl. 2019;54:798–809.
20. Cornelissen JJ, Gratwohl A, Schlenk RF, Sierra J, Bornhäuser M, Juliusson G, et al. The European LeukemiaNet AML Working Party consensus statement on allogeneic HSCT for patients with AML in remission: an integrated-risk adapted approach. Nat Rev Clin Oncol. 2012;9:579–90.
21. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Buchner T, et al. Diagnosis and management of AML in adults:

2017 ELN recommendations from an international expert panel. Blood. 2017;129:424–47.
22. Getta BM, Devlin SM, Levine RL, Arcila ME, Mohanty AS, Zehir A, et al. Multicolor flow cytometry and multigene next-generation sequencing are complementary and highly predictive for relapse in acute myeloid leukemia after allogeneic transplantation. Biol Blood Marrow Transpl. 2017;23:1064–71.
23. Jongen-Lavrencic M, Grob T, Hanekamp D, Kavelaars FG, Al Hinai A, Zeilemaker A, et al. Molecular minimal residual disease in acute myeloid leukemia. N Engl J Med. 2018;378:1189–99.
24. Altman DG, Royston P. The cost of dichotomising continuous variables. Brit Med J. 2006;332:1080.
25. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. Bone Marrow Transplant. 2019. https://doi.org/10.1038/s41409-019-0679-x [Epub ahead of print].
26. Estey E, Gale RP. How good are we at predicting the fate of someone with acute myeloid leukaemia? Leukemia. 2017;31:1255–8.
27. van Dongen JJM, van der Velden VHJ, Brüggemann M, Orfao A. Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood. 2015;125:3996–4009.
28. Theunissen P, Mejstrikova E, Sedek L, van der Sluijs-Gelling AJ, Gaipa G, Bartels M, et al. Standardized flow cytometry for highly sensitive MRD measurements in B-cell acute lymphoblastic leukemia. Blood. 2017;129:347–57.
29. Pfeifer H, Cazzaniga G, van der Velden VHJ, Cayuela JM, Schäfer B, Spinelli O, et al. Standardisation and consensus guidelines for minimal residual disease assessment in Philadelphia-positive acute lymphoblastic leukemia (Ph + ALL) by real-time quantitative reverse transcriptase PCR of e1a2 BCR-ABL1. Leukemia. 2019;33:1910–22.
30. Flores-Montero J, Sanoja-Flores L, Paiva B, Puig N, García-Sánchez O, Böttcher S, et al. Next generation flow for highly sensitive and standardized detection of minimal residual disease in multiple myeloma. Leukemia. 2017;31:2094–103.
31. Houshmand M, Simonetti G, Circosta P, Gaidano V, Cignetti A, Martinelli G, et al. Chronic myeloid leukemia stem cells. Leukemia. 2019;33:1543–56.
32. Pott C, Brüggemann M, Ritgen M, van der Velden VHJ, van Dongen JJM, Kneba M. MRD detection in B-cell non-Hodgkin lymphomas using Ig gene rearrangements and chromosomal translocations as targets for real-time quantitative PCR. Methods Mol Biol. 2019;1956:199–228.
33. Böttcher S. Flow cytometric MRD detection in selected mature B-cell malignancies. Methods Mol Biol. 2019;1956:157–97.
34. Hodes A, Calvo KR, Dulau A, Maric I, Sun J, Braylan R. The challenging task of enumerating blasts in the bone marrow. Semin Hematol. 2019;56:58–64.
35. Petersdorf SH, Kopecky KJ, Slovak M, Willman C, Nevill T, Brandwein J, et al. A phase 3 study of gemtuzumab ozogamicin during induction and postconsolidation therapy in younger patients with acute myeloid leukemia. Blood. 2013;121:4854–60.
36. Othus M, Wood BL, Stirewalt DL, Estey EH, Petersdorf SH, Appelbaum FR, et al. Effect of measurable ('minimal') residual disease (MRD) information on prediction of relapse and survival in adult acute myeloid leukemia. Leukemia. 2016;30:2080–3.
37. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med. 2013;32:5381–97.
38. Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. Biometrics. 2012;68:388–96.
39. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004;159:882–90.
40. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35:4124–35.
41. Fauci AS. Human bone marrow lymphocytes. I. Distribution of lymphocyte subpopulations in the bone marrow of normal individuals. J Clin Investig. 1975;56:98–110.
42. Gale RP, Opelz G, Kiuchi OM, Golde DW. Thymus-dependent lymphocytes in human bone marrow. J Clin Investig. 1975;56:1491–8.
43. Martens ACM, Schultz FW, Hagenbeek A. Nonhomogeneous distribution of leukemia in the bone marrow during minimal residual disease. Blood. 1987;70:1073–8.
44. Butturini A, Klein J, Gale RP. Modeling minimal residual disease (MRD)-testing. Leuk Res. 2003;27:293–300.
45. Soni PD, Hartman HE, Dess RT, Abugharib A, Allen SG, Feng FY, et al. Comparison of population-based observational studies with randomized trials in oncology. J Clin Oncol. 2019;37:1209–16.
46. Terwijn M, van Putten WL, Kelder A, van der Velden VH, Brooimans RA, Pabst T, 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. 2013;31:3889–97.
47. Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R, et al. Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol. 2013;14:199–209.
48. Vora A, Goulden N, Mitchell C, Hancock J, Hough R, Rowntree C, et al. Augmented post-remission therapy for a minimal residual disease-defined high-risk subgroup of children and young people with clinical standard-risk and intermediate-risk acute lymphoblastic leukaemia (UKALL 2003): a randomised controlled trial. Lancet Oncol. 2014;15:809–18.
49. Schrappe M, Bleckmann K, Zimmermann M, Biondi A, Möricke A, Locatelli F, et al. Reduced-intensity delayed intensification in standard-risk pediatric acute lymphoblastic leukemia defined by undetectable minimal residual disease: results of an international randomized trial (AIEOP-BFM ALL 2000). J Clin Oncol. 2018;36:244–53.
50. Hourigan C, Dillon L, Logan B, Scott B, Ghannam J, Gui G, et al. Impact of conditioning intensity of allogeneic transplantation for acute myeloid leukemia with genomic evidence of residual disease. In: 24th Congress of the European Hematology Association. Amsterdam, the Netherlands; 2019.
51. Rubnitz JE, Hijiya N, Zhou Y, Hancock ML, Rivera GK, Pui CH. Lack of benefit of early detection of relapse after completion of therapy for acute lymphoblastic leukemia. Pediatr Blood Cancer. 2005;44:138–41.
52. Wille-Jørgensen P, Syk I, Smedh K, Laurberg S, Nielsen DT, Petersen SH, et al. Effect of more vs less frequent follow-up testing on overall and colorectal cancer-specific mortality in patients with stage II or III colorectal cancer: the COLOFOL randomized clinical trial. JAMA. 2018;319:2095–103.
53. Snyder RA, Hu CY, Cuddy A, Francescatti AB, Schumacher JR, Van Loon K, et al. Association between intensity of posttreatment surveillance testing and detection of recurrence in patients with colorectal cancer. JAMA. 2018;319:2104–15.