EUROPEAN HEMATOLOGY ASSOCIATION
haematologica
Journal of the European Hematology Association

The eGVHD App has the potential to improve the accuracy of graft versus host disease assessment: a multicenter randomized controlled trial

by Helene M. Schoemans, Kathy Goris, Raf Van Durm, Steffen Fieuws, Sabina De Geest, Steven Z. Pavletic, Annie Im, Daniel Wolff, Stephanie J. Lee, Hildegard Greinix, Rafael F. Duarte, Xavier Poiré, Dominik Selleslag, Philippe Lewalle, Tessa Kerre, Carlos Graux, Frédéric Baron, Johan A. Maertens, and Fabienne Dobbels

Haematologica 2018 [Epub ahead of print]

Citation: Helene M. Schoemans, Kathy Goris, Raf Van Durm, Steffen Fieuws, Sabina De Geest, Steven Z. Pavletic, Annie Im, Daniel Wolff, Stephanie J. Lee, Hildegard Greinix, Rafael F. Duarte, Xavier Poiré, Dominik Selleslag, Philippe Lewalle, Tessa Kerre, Carlos Graux, Frédéric Baron, Johan A. Maertens, and Fabienne Dobbel. The eGVHD App has the potential to improve the accuracy of graft versus host disease assessment: a multicenter randomized.
Haematologica. 2018; 103:xxx
doi:10.3324/haematol.2018.190777

The eGVHD App has the potential to improve the accuracy of graft versus host disease assessment: a multicenter randomized controlled trial

Helene M. Schoemans[1,2], Kathy Goris[1], Raf Van Durm[3], Steffen Fieuws[4], Sabina De Geest[2,5], Steven Z. Pavletic[6], Annie Im[7], Daniel Wolff[8], Stephanie J. Lee[9], Hildegard Greinix[10], Rafael F. Duarte, MD, PhD[11], Xavier Poiré[12], Dominik Selleslag[13], Philippe Lewalle[14], Tessa Kerre[15], Carlos Graux[16], Frédéric Baron[17], Johan A. Maertens[1] and Fabienne Dobbels, PhD[2], *on behalf of the EBMT Transplantation Complications Working party.*

Affiliations

(1) Department of Hematology, University Hospitals Leuven and KU Leuven, Leuven, Belgium
(2) Academic Centre for Nursing and Midwifery, KU Leuven, Leuven, Belgium
(3) IT Department, University Hospitals Leuven, KU Leuven, Leuven, Belgium,
(4) I-Biostat, KU Leuven – University of Leuven & Universiteit Hasselt, Leuven, Belgium
(5) Institute of Nursing Science, Department Public Health, University of Basel, Basel, Switzerland
(6) Experimental Transplantation and Immunology Branch, Center for Cancer Research (CCR), National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD
(7) University of Pittsburgh Medical Center, Pittsburgh, PA
(8) Department of Hematology and Clinical Oncology, University of Regensburg, Regensburg, Germany
(9) Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA,
(10) Division of Hematology, Medical University of Graz, Graz, Austria,
(11) ICO/Hospital Duran I Reynals, Hospitalet De Llobregat, Spain,
(12) Cliniques Universitaires Saint-Luc, Brussels, Belgium,
(13) Dept of Hematology, AZ Sint Jan Brugge, Brugge, Belgium,
(14) Institut Jules Bordet-Université Libre de Bruxelles, Brussels, Belgium,
(15) Hematology and Stem Cell Transplantation, Ghent University Hospital, Ghent, Belgium,
(16) Université Catholique de Louvain, CHU UCL Namur (Godinne site), Yvoir, Belgium,
(17) Hematology, University of Liège, GIGA-I3, Liège, Belgium

Authorship

HMS coordinated the project, performed the research and wrote the manuscript. All authors (HMS, KG, RVD, SF, SDG, SZP, AI, DW, SJL, HG, RFD, XP, DS, PL, TK, CG, FB, JAM and FD) participated in the research and reviewed the manuscript. KG was responsible for the study organization and data management. RVD designed the App. SZP, AI, DW, SJL and HG served as GvHD experts. XP, DS, PL, TK, CG, FB, and JAM organized the local workshops. HMS, KG, SF, SDG, JAM and FD designed the research and performed the statistical analysis. All authors reviewed the final version and approved submission.

Conflict of interest

The authors declare that they have no competing financial interests related to the work described. The material is original research. It has been as accepted for presentation as a poster at the 2018 BMT Tandem Meetings and as an oral abstract at EBMT 2018. It has not been previously published and it has not been submitted for publication elsewhere.

Keywords

GvHD, HCT, HSCT, BMT, clinical practice, accuracy, usability, eHealth, mobile application.

Running Title

Impact of the eGVHD App on GvHD assessment

Corresponding Author

Hélène Schoemans

Department of Hematology, University Hospitals Leuven and KU Leuven, Leuven, Belgium

49 Herestraat, B 3000 Leuven, Belgium

Tel +32 16 346889

Fax +32 16 346881

helene.schoemans@uzleuven.be

ABSTRACT :

Graft-versus-host disease assessment has been shown to be a challenge for healthcare professionals, leading to the development of the eGVHD App (www.uzleuven.be/egvhd). In this study, we formally evaluated the accuracy of using the App compared to traditional assessment methods to assess graft-versus-host disease. Our national multicenter randomized controlled trial involved seven Belgian transplantation centers and 78 healthcare professionals selected using a two-stage convenience sampling approach between January and April 2017. Using a 1:1 randomization stratified by profession, healthcare professionals were assigned to use either the App ("APP") or their usual graft-versus-host disease assessment aids ("No APP") to assess the diagnosis and severity score of ten expert-validated clinical vignettes. Our main outcome measure was the difference in accuracy for graft-versus-host disease severity scoring between both groups. The odds of being correct were 6.14 (95% CI: 2.83-13.34) and 6.29 (95% CI: 4.32-9.15) times higher in favor of the "APP" group for diagnosis and scoring, respectively ($p<0.001$). App-assisted graft-versus-host disease severity scoring was significantly superior for both acute and chronic graft-versus-host disease, with an Odds Ratio of 17.89 and 4.34 respectively ($p<0.001$) and showed a significantly increased inter-observer agreement compared to standard practice. Despite a mean increase of 24 minutes (95% CI: 20.45-26.97) in time needed to score the whole graft-versus-host disease test package in the "APP" group ($p<0.001$), usability feedback was positive. The eGVHD App showed superior graft-versus-host disease assessment accuracy compared to standard practice and has the potential to improve the quality of outcome data registration in allogeneic stem cell transplantation.

Abstract word count: 250

INTRODUCTION:

Graft-versus-host Disease (GvHD) refers to the reaction of the transplanted immune system against the recipient's tissues. This pleiotropic disease affects up to half of patients after allogeneic hematopoietic stem cell transplantation (HCT) and can damage any organ system to various degrees. It is by far the most debilitating complication of HCT, considering its major impact on morbidity and mortality[1].

Yet, because of the lack of widely available GvHD biomarkers, the assessment of the presence and severity of GvHD still relies mainly on the clinical evaluation of multiple organs according to a relatively complex algorithm. Moreover, the recommendations underlying this evaluation are plethoric and sometimes even contradictory, potentially leading to confusion in the HCT community[1]. In fact, it has been repeatedly shown that many HCT professionals have difficulties with the correct implementation of GvHD assessment, as demonstrated by a low observed accuracy in GvHD assessment[2-5] and a slow uptake of the most up-to-date guidelines[5-7].

The eGVHD App is an electronic tool that we developed in collaboration with the EBMT Transplantation Complications Working Party and the National Institutes of Health (NIH) to assist healthcare professionals with their GvHD assessment[4]. This tool is a web application, available on mobile devices and desktop computers (see www.uzleuven.be/egvhd for a complete list of the App's characteristics). It allows intuitive and user-friendly access to the most recent international consensus guidelines and assists the user by automatically executing the required algorithm to calculate severity of GvHD, once the relevant clinical characteristics have been entered.

Pilot testing was promising, suggesting improved GvHD assessment and good usability[4, 5]. Therefore, the primary aim of the present study was to compare the **accuracy of the severity score** of validated GvHD case-vignettes performed by healthcare professionals using the "eGVHD app" ("APP" group) with standard practice ("No APP" group). Secondary aims were to understand the characteristics that might affect the difference in accuracy between both groups and to compare the **inter-observer variability** in GvHD scoring results as well as the **time** needed to perform the GvHD evaluation of the full test package in both groups. We also assessed current **practice patterns** in GvHD assessment for all participants and **post-test user satisfaction and experience** in the "APP" group, to allow further improvement of the tool's usability. To evaluate the **generalizability** of the tool, we tested the eGVHD App in a variety of settings and with a wide range of healthcare practitioners with different professional backgrounds.

We hypothesized that the eGVHD App would improve GvHD assessment by improving the accuracy of GvHD severity scoring by healthcare professionals and reducing inter-rater variability in scoring results, without increasing the time required to assess GvHD.

METHODS:

Design:

This study used a hybrid design (**Figure 1**). The first part of the study consisted of a two-group multicenter randomized controlled trial assigning healthcare professionals 1:1 to an intervention group ("**APP"**) or a control group ("**No APP"**) to evaluate the accuracy of GvHD assessment. The second part of the study was observational and described current practice patterns in GvHD assessment ("Survey 1") and usability aspects linked to the use of the App ("Survey 2").

Sample and setting:

All Belgian hospitals performing allogeneic HCT were invited to participate (**Supplementary Table 1**) to optimize sample size and generalizability. Centers were selected on their willingness to organize a GvHD workshop on their own premises within the allocated timeframe (January to April 2017). Healthcare professionals employed or studying at each participating hospital were recruited by convenience sampling. They were included provided they attended the workshop (**see supplemental methods for workshop details**) and could recall having performed at least one GvHD evaluation in the past 12 months.

Data collection points, randomization procedure and blinding: see supplemental methods

Outcome measures:

The primary aim was to assess the difference in accuracy for GvHD severity scoring between the "APP" and "No APP" groups (**see supplemental methods for the planned sub-analyses**).

Variables and measurements:

Demographics and practice patterns in GvHD assessment:

A self-report questionnaire ("Survey 1") captured participant characteristics (**Table 1**) as well as practice pattern in GvHD assessment and pre-test technology access & acceptance data (**Table 2**) at baseline.

Accuracy of GvHD assessment:

Participants were required to diagnose and score a package of ten randomly ordered GvHD clinical vignettes based on real-life clinical cases (**see Supplemental Methods and Supplementary Table 2**) according to the most up to date international guidelines[1]. Four aGvHD vignettes covered the two types of aGvHD diagnosis ('classic aGvHD' and 'late aGvHD', two vignettes each) and the four aGvHD overall severity stages (I-IV, one vignette per stage), according to the MAGIC criteria[8]. Six cGvHD vignettes covered the two cGvHD diagnoses ('overlap cGvHD' and 'classic cGvHD', two and four vignettes respectively) and the three severity grades of the NIH 2014 criteria[9] (two vignettes per severity level, i.e. mild, moderate and severe). Answers were given by participants using a multiple choice form offering the following mutually exclusive options for diagnosis ('classic aGvHD', 'late aGvHD', 'overlap cGvHD' or 'classic cGvHD') and scoring ('grade I', 'grade II', 'grade III', 'grade IV', 'Mild', 'Moderate' or 'Severe') respectively.

The individual answer of each participant was compared to the gold standard (**see Supplemental Methods**) and scored as 'correct' (if the answer corresponded exactly to the expert evaluation) or 'incorrect' (for any other answer, including missing answers) for diagnosis and severity scoring respectively (**Supplementary Table 3**). The total number of correctly evaluated vignettes for the whole GvHD test package was also recorded per individual (score ranging from 0 to 10 correct answers), for diagnosis and scoring separately. The time needed to complete the full GvHD test package was recorded for each participant individually by study staff.

Control Group:

Participants randomized to standard practice ("**No APP**" **control group**) were allowed to use any of their usual methods to assess GvHD: their own knowledge, 'fast facts' sheets, scoring sheets, standard operating procedures, copies of original guideline publications, or any other chosen resource.

Intervention Group:

Participants randomized to the "**APP**" group received the eGVHD App as a stand-alone GvHD assessment aid.

Post-test user satisfaction and experience:

Post-test user satisfaction and experience was recorded in "APP" users only by "Survey 2" using a semi-structured self-report questionnaire, and two validated instruments, the "perceived usefulness" subscale of the technology acceptance model (TAM) and the Post-Study System Usability Questionnaire (PSSUQ), as described previously[4] (**see supplemental methods and Supplementary Table 4 for details**).

Statistical analysis: see supplemental methods

RESULTS

Seven out of the eleven Belgian allogeneic HCT centers participated in the study (response rate 64%). They were essentially academic centers, covering together more than 80% of the Belgian allogenic transplantation activity (**Supplementary Table 1**).

A total of 103 individuals participated in the workshops (**Figure 2**). Seventy-eight professionals met the inclusion criteria and were randomized. One participant dropped-out, due to a medical emergency in the clinic, hence data from 77 professionals were available for analysis: 37 in the "APP" Group and 40 in the "No APP" group. There was a median of 8 participants per center (range: 7-20, **Supplementary Table 1**). Professional characteristics were similar in both groups (**Table 1**). The majority of participants were medical doctors (75%), female (64%), and had a median age of 39 years (IQR: 20, range 22-62). Professionals reported a median experience in allogeneic HCT of 6 years (IQR: 11, range 0-32), and evaluated a median of one allogeneic HCT patient for GvHD per week (IQR: 5, range 0-30). The majority of healthcare professionals reported having expertise in adult patient care. Self-reported proficiency in English was high with a median of 7 (IQR: 1; range: 2-10).

Pre-test user current standard practice and technology access/acceptance

The Glucksberg[10] and the NIH 2014 criteria[9] were the most frequently referenced GvHD assessment guidelines being used in clinical practice as reported by healthcare professionals (**Table 2**). Most professionals reported basing their usual GvHD evaluation on their own knowledge (n= 44, 57%), the NIH 2014 GvHD evaluation sheet[9] (n=17, 22%) and/or a self-designed scoring paper document (n= 16, 21%). The use of standard criteria to assess GvHD was reported as important (median score of 7 on a Likert scale of one to ten, IQR: 4, range 1-10), but performed with a relatively low level of confidence (median score of 5 on a Likert scale of one to ten, IQR : 4, range 1-9). The top four GvHD assessment problems spontaneously reported were: lack of knowledge or experience (n=23), time constraints (n= 16), lack of data in the medical files (n=7) and the complexity of the guidelines (n=5).

During the workshop, the "No APP" group planned to rely essentially on their own knowledge (n=24, 62%), the NIH 2014 GvHD evaluation sheet[9] (n=9, 23%), the NIH 2005 GvHD evaluation sheet[11] (n=6, 15%), a self-designed scoring document (n=6, 15%), and/or other methods (n=7, 18%) (**Table 2**).

Accuracy of GvHD assessment

The total number of correctly evaluated clinical vignettes was higher in the "APP" group compared to the "No APP" group (**Table 3**). More specifically, participants in the "APP" group had a median of 10 correct answers for diagnosis (IQR 1; range 5-10), compared to a median of 6.5 (IQR 3; range 2-9) in the "No APP" group for the whole GvHD test package (the maximum obtainable score was 10). For severity assessment, the "APP" group scored a median of 9 vignettes correctly (IQR 2; range 2-10) compared to a median of 4.5 (IQR 3; range 1-7) in the "No APP" group. Individual results for each vignette are shown in **Supplementary Table 3**. As a result, the odds of being correct were 6.14 (95% CI 2.83-13.34) and 6.29 (95% CI 4.32-9.15) times higher in favor of the "APP" group for diagnosis and scoring respectively ($p<0.001$).

All pre-specified sub-analyses were performed as planned. The GvHD assessment of the "APP" group remained superior for both acute and chronic GvHD separately (with a significantly stronger effect in acute GvHD (OR =17.89, 95% CI 8.47-37.79) compared to chronic GvHD (OR=4.34, 95% CI 2.79-6.74, $p<0.001$), and for all levels of severity scoring, except for aGvHD grade I. The effect of the App was more apparent for higher levels of severity ($p=0.034$) for both aGvHD and cGvHD. The strength of the effect did not significantly depend on center (**Supplementary Figure 1**) or professional background (**Supplementary Figure 2**). Similarly, neither the age of user (**Supplementary Figure 3**), the number of GvHD patients seen per week (**Supplementary Figure 4**) or self-reported comfort with using GvHD guidelines (**Supplementary Figure 5**) seemed to mitigate the superior performance of the "APP" group.

Agreement between participant results and the expert gold standard diagnosis and severity scoring are highlighted in the diagonal of **Table 4**, showing the superior performance of the "APP" group. For diagnosis, the most consistent errors of the "No APP" group were seen for case-vignettes relating to 'Overlap cGvHD' and 'Late aGvHD', which both tended to be confused with 'Classic cGvHD'. The highest discrepancies between the "No APP" group and expert acute GvHD severity scoring results were seen in 'grade II' (which tended to be graded according to the cGvHD criteria) and 'grade IV' aGvHD (which was essentially mistaken for 'grade III'). Inconsistencies in chronic GvHD severity scoring were seen across all grades. The most frequent error in the "APP" group was a slight overestimation of the

cGvHD grade (overestimation n=34, 15%; underestimation n=20, 9%; missing/other n=4, 2%) without any misclassification, whereas the "No APP" group tended to evaluate cGvHD severity erroneously according to the aGvHD criteria (n=62, 25%), without bias for severity (overestimation n=36, 14%, underestimation n=36, 15%, missing/other n=7, 3%).

Consequently, inter-observer agreement of the severity score was higher in the "APP" group compared to standard practice: the probability that two HCT professionals agreed on the GvHD score equaled 0.73 and 0.56 in the "App" and No APP" group, respectively. The chance-corrected agreement was significantly higher in the "APP" group (K_{BP}= 0.46, 95% CI: 0.23-0.68) compared to the "No APP" group (K_{BP} =0.12, 95% CI: 0.03-0.21) (p=0.003).

The time needed to complete the total test package was significantly higher in the "APP" group compared to the standard practice group, with a mean time of 48.84 minutes to complete all ten clinical vignettes in the "APP" group versus 25.27 minutes in the "No APP" group (p<0.001) (**Table 3**).

Post-Test User satisfaction and experience

No major technical issues were identified. Both "perceived usefulness" and "system usability" were considered to be good as shown in **supplementary Table 4**. Users reported being likely to use the eGVHD App in their daily practice and did not experience any issues with using the App in English. Spontaneously reported positive aspects of the eGVHD App were its clarity, ease of use and its systematic approach. Users suggested some potential improvements, such as decreasing its time-consuming components, reducing the number of evaluated items and clarifying some specific terms in more detail.

DISCUSSION

Several groups have recently advocated the use of electronic tools to improve GvHD assessment, albeit without providing formal proof of their efficacy[1, 4, 12-14]. In this rigorous multi-center randomized trial, we unequivocally demonstrate that the accuracy of GvHD assessment of clinical vignettes by healthcare professionals is significantly higher when using the eGVHD App compared to standard practice. This effect was seen for both acute and chronic GvHD, across all severity levels (except for aGvHD grade I), all degrees of experience and professional backgrounds, without any evidence for center effect.

In this study, participants in the control group were allowed to use any method of their choice to support their GvHD assessment, except for using the eGVHD App. Yet GvHD assessment results in the APP group, were strikingly better. We believe that the superior performance of the App users could be due to a

number of factors. First, App users were provided with the most up-to-date guidelines[1], without having to look them up actively. Second, similar to using comprehensive paper data collection forms, they were encouraged to work in a systematic fashion: they had to evaluate every possible aspect of acute or chronic GvHD (to avoid overlooking less intuitive aspects of the disease) in order to select the appropriate scoring system and come to the correct severity evaluation result. Finally, the digital interface also offered users a number of advantages such as the presence of pictures and definitions to support recognition of GvHD-related features, the use 'skip-logic' principles (which allows healthcare professionals to avoid wasting time on filling in information with no direct impact on diagnosis or severity scoring), the automatic computation of the resulting score and the option of generating a report.

We have to acknowledge that this superior performance was achieved at the cost of a significant increase in time needed to score clinical vignettes, with an excess of about 24 minutes to score the ten clinical vignettes compared to using standard methods. This was partially due to the fact that "APP" users needed to get used to a tool they had never worked with before. Yet, healthcare professionals remained open to the use of eHealth technology, both before and after actually using the App. The eGVHD App showed excellent usability, as no major technical issues were noted and user feedback was widely positive, suggesting a potential for optimal dissemination and uptake in the HCT community. Furthermore, in the event where the App-computed scores would be directly transferred into the electronic medical record (eHR), the additional time spent inputting data into the App would be rewarded with potentially less time charting and more accurate data collection. However, this integration also supposes a number of basic pre-requisites, which still need to be developed: data cleaning methods to ensure the quality of data entry, the possibility of cross-talk between the eGVHD App and the different eHR systems, the reliability, privacy and safety of data transfer and the option of identifying the individual who performed the data input.

Consistent with prior literature, our practice pattern survey showed the lack of consensus in the HCT community as to which set of international recommendations should be used to assess GvHD, and confirmed numerous barriers to their successful dissemination and implementation[5-7]. The lack of consensus and knowledge of the most recent guidelines was maybe due to the low number of HCT patients seen per week and probably partly explains the lower results obtained by the group using traditional methods. However, this also highlights the need to standardize GvHD evaluation within the HCT community, as recently advocated by a panel of GvHD experts[1]. It is precisely in this context of lack of confidence and expertise in GvHD assessment that e-Tools, such as the eGVHD App, have the potential to increase the quality of data collection by allowing for an easy, reliable, user-friendly and intuitive access to the most up-to-date guidelines to any healthcare professional. Regrettably, we were

unable to test the effect of the App specifically in smaller Belgian centers, as they declined participation to this study. We are therefore unable to speculate on the generalizability of this tool in centers with lower transplantation volumes.

The limited number of vignettes also makes it challenging to make any meaningful conclusions on specific subgroups or at the organ level. The significant difference in improved accuracy for aGvHD scoring compared to cGvHD scoring is probably simply due to the fact that each of the four aGvHD severity levels was evaluated by a single clinical vignette (instead of two per severity level for cGvHD). For instance, in the 'late acute GvHD grade II' clinical vignette, the largely incorrect final severity evaluation reported by the "No APP" group was partially conditioned by the fact that the distinction between acute and chronic GvHD had not been made in the first place. Moreover, the MAGIC criteria were not the standard reference for aGvHD for the majority of the participants, which could explain the exceptionally poor results for the grade IV aGvHD vignette when evaluated by the "No APP" group.

The limited number of observations also restrict our ability to make any conclusions on the potential impact of using the App in the clinical setting to decide upon starting treatment, as the threshold to start therapy is linked to much broader categories than the ones described above (typically, any grade above or equal to 'aGvHD grade II' or 'cGvHD moderate' would qualify for treatment, depending on the general health status of the patient[15-17]). Treatment adaptions rely also on specify response criteria[18, 19], which were not investigated in this project. Future work therefore needs to evaluate the use and impact of the eGVHD App in clinical encounters. This will also allow the evaluation of the App in situations where the patient *does not* present with GvHD, considering that the test package evaluated here only evaluated the tool in the context of GvHD-afflicted patients, precluding the evaluation of detection measures such as predictive values, sensitivity and specificity.

Further limitations of this study are the lack of repeated measures and the unnatural setting of clinical vignettes, which are unable to perfectly mirror the wide variations in GvHD presentation in real life and their relative incidence. This particular experimental design was chosen to simplify logistics, optimize healthcare professional participation, avoid patient bother and keep respondent burden to a minimum. It also allowed for multiple experts to validate the GvHD assessment. Such an expert consensus is rarely obtained in clinical practice, but was considered to be the best gold standard available to date to serve as reference for the accurate scoring during GvHD assessment.

So, it remains to be determined whether the App will also improve accuracy when being used in real life circumstances. Yet, even in this artificial setting, the low spontaneous GvHD scoring accuracy obtained in this evaluation with traditional methods (obtaining a median of 4.5 correctly scored vignettes out of a maximum of ten) is in line with the results of a previous validation study done in a more real-world setting. This study included actual patient examinations and showed that only 50% to 75% of freshly trained clinicians actually agreed with experts on the overall severity score of the evaluated chronic GvHD patients [6]. Mitchell and colleagues concluded that a single training session was insufficient to achieve consistently acceptable inter-rater agreement between novice healthcare practitioners and GvHD experts. Clinical training in GvHD physical exams may thus be necessarily to achieve reproducible severity assessment with high inter-rater reliability in practice. By ensuring the systematic assessment of all organs potentially affected by GvHD, the App can also serve as a training tool, aiming at making healthcare professionals ultimately independent of technological assistance.

The eGVHD App is currently limited to a calculator function that evaluates the patient at a single point in time. Expanding on our promising accuracy results and user-feedback, future plans include the development of a module to perform longitudinal patient evaluations (with an integrated disease response evaluation according to international criteria[18, 19]) and a module to capture patient-reported GvHD evaluation based on the Lee symptom scale[20]. These added functionalities will dramatically increase the clinical usefulness of the tool in following patients over time.

However, a challenging issue with eHealth tools is how to approach their constant and rapid change over time. This evolution is driven by evolving clinical practices, user feedback and updates in computer programs and/or operating systems. The results reported in this study, for instance, have been obtained with a version of the eGVHD app, which has already become obsolete, as a new version (using additional skip-logic features) has been developed to address the rightful criticism about the time-consuming aspect of use. The constant evolution of the virtual world is a challenge in the current context of European regulation (EU Directive 93/42/EEC MEDDEV 2. 4/1 Rev. 9 June 2010), which requires eHealth applications to be formally validated by a tedious quality insurance process at every adaptation of the tool. This is not practically feasible in real life and probably more often than not, unnecessary. Health regulations agencies will need to adjust their requirements in the near future, to allow for this dynamic progress of the cyber world, even for healthcare applications. This is, in fact, probably one of the most challenging aspects of integrating eTools in modern models of care[21].

Compared to other smaller scaled initiatives, which have shown successful implementation of eHealth technologies in local electronic medical record systems [14] or specific research programs [12, 13] to assess GvHD, the eGVHD App is now ubiquitously available (www.uzleuven.be/egvhd) for all healthcare professionals who wish to get bedside user-friendly assistance in their GvHD assessment, to improve their expertise and/or the uniformity of their GvHD data collection, both in daily practice and in clinical trials. Further validation regarding its usefulness and scalability will therefore be able to rely on the analysis of the real-life data generated by downloads and feedback from users, based on implementation research principles. If results are convincing, next steps could include the direct integration of eGVHD App-generated data in larger registry databases and electronic medical record systems to circumvent the need to produce separate reports and repeat data entry. Such developments will require further reflections on how to achieve optimal control of the quality of the entered data and guarantee its privacy protection according to local laws.

In conclusion, the eGVHD App shows superior accuracy for the GvHD assessment of clinical vignettes compared to usual care and has therefore the potential to improve the quality of GvHD data in clinical research and practice. In the era of electronic medical files, 'big data' and increased connectivity, e-Tools are likely to become widespread in our daily practice and could even gradually turn the patient-himself into his own data-manager and most involved advocate. Only time and continuous research will tell whether such tools can be effectively used in clinical practice and whether healthcare professionals are ready to accept IT assistance to solve some of their practical issues.

Article word count: 3941

Acknowledgements:

The authors would like to thank all of the participating hospitals for their collaboration and enthusiasm in validating the eGVHD App. We are also very grateful for the financial support of SOFHEA vzw (Sociaal Fonds voor Hematologische Aandoeningen) for this project.

REFERENCES:

1. Schoemans HM, Lee SJ, Ferrara JL, et al. EBMT-NIH-CIBMTR Task Force position statement on standardized terminology & guidance for graft-versus-host disease assessment. Bone Marrow Transplant. 2018 Jun 5. [Epub ahead of print]
2. Carpenter PA, Logan BR, Lee SJ, et al. Prednisone (PDN)/Sirolimus (SRL) Compared to PDN/SRL/Calcineurin Inhibitor (CNI) as Treatment for Chronic Graft-Versus-Host-Disease (cGVHD): A Randomized Phase II Study from the Blood and Marrow Transplant Clinical Trials Network. Biol Blood Marrow Transplant. 2016;22(3):S50-S52.
3. Weisdorf DJ, Hurd D, Carter S, et al. Prospective grading of graft-versus-host disease after unrelated donor marrow transplantation: a grading algorithm versus blinded expert panel review. Biol Blood Marrow Transplant. 2003;9(8):512-518.
4. Schoemans H, Goris K, Durm RV, et al. Development, preliminary usability and accuracy testing of the EBMT 'eGVHD App' to support GvHD assessment according to NIH criteria-a proof of concept. Bone Marrow Transplant. 2016;51(8):1062-1065.
5. Schoemans HM, Goris K, Van Durm R, et al. Accuracy and usability of the eGVHD app in assessing the severity of graft-versus-host disease at the 2017 EBMT annual congress. Bone Marrow Transplant. 2018;53(4):490-494.
6. Mitchell SA, Jacobsohn D, Thormann Powers KE, et al. A multicenter pilot evaluation of the National Institutes of Health chronic graft-versus-host disease (cGVHD) therapeutic response measures: feasibility, interrater reliability, and minimum detectable change. Biol Blood Marrow Transplant. 2011;17(11):1619-1629.
7. Duarte RF, Greinix H, Rabin B, et al. Uptake and use of recommendations for the diagnosis, severity scoring and management of chronic GVHD: an international survey of the EBMT-NCI Chronic GVHD Task Force. Bone Marrow Transplant. 2014;49(1):49-54.
8. Harris AC, Young R, Devine S, et al. International, Multicenter Standardization of Acute Graft-versus-Host Disease Clinical Data Collection: A Report from the Mount Sinai Acute GVHD International Consortium. Biol Blood Marrow Transplant. 2016;22(1):4-10.
9. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. The 2014 Diagnosis and Staging Working Group report. Biol Blood Marrow Transplant. 2015;21(3):389-401.
10. Glucksberg H, Storb R, Fefer A, et al. Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors. Transplantation. 1974;18(4):295-304.
11. Filipovich AH, Weisdorf D, Pavletic S, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. Diagnosis and staging working group report. Biol Blood Marrow Transplant. 2005;11(12):945-956.
12. Levine JE, Hogan WJ, Harris AC, et al. Improved accuracy of acute graft-versus-host disease staging among multiple centers. Best Pract Res Clin Haematol. 2014;27(3-4):283-287.
13. Mancini G, Frulla R, Vico M, e al. A new software for evaluating scoring and response in cGVHD according to the new NIH criteria. Bone Marrow Transplant. 2016;51(Issue S1):S183.
14. Dierov Djamilia CC, Fatmi S, Mosesso K, et al . Establishing a standardized system to capture chronic graft-versus-host disease (GVHD) data in accordance to the national institutes (NIH) consensus criteria. Bone Marrow Transplant. 2017;52 (Suppl 1):S102 (abstract O157).
15. Deeg HJ. How I treat refractory acute GVHD. Blood. 2007;109(10):4119-4126.
16. Martin PJ, Schoch G, Fisher L, et al. A retrospective analysis of therapy for acute graft-versus-host disease: initial treatment. Blood. 1990;76(8):1464-1472.

17. Wolff D, Gerbitz A, Ayuk F, et al. Consensus conference on clinical practice in chronic graft-versus-host disease (GVHD): first-line and topical treatment of chronic GVHD. Biol Blood Marrow Transplant. 2010;16(12):1611-1628.
18. Lee SJ, Wolff D, Kitko C, et al. Measuring therapeutic response in chronic graft-versus-host disease. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. The 2014 Response Criteria Working Group report. Biol Blood Marrow Transplant. 2015;21(6):984-999.
19. MacMillan ML, Robin M, Harris AC, et al. A Refined Risk Score for Acute Graft-versus-Host Disease that Predicts Response to Initial Therapy, Survival, and Transplant-Related Mortality. Biol Blood Marrow Transplant. 2015;21(4):761-767.
20. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2002;8(8):444-452.
21. Tuckson RV, Edmunds M, Hodgkins ML. Telehealth. N Engl J Med. 2017;377(16):1585-1592.

Tables

Table 1: Characteristics of workshop participants

Professional Background	Whole group (n= 77)	APP (n=37)	No APP (n=40)
Senior physicians - n (%)	37 (48%)	18 (49%)	19 (48%)
Junior physicians - n (%)	21 (27%)	10 (27%)	11 (27%)
Data managers - n (%)	15 (19%)	7 (19%)	8 (20%)
Others - n (%)	4 (5%)	2 (5%)*	2 (5%)**
Demographics			
Gender - n (%)	28 males (36%) 49 females (64%)	13 males (35%) 24 females (65%)	15 males (37%) 25 females (62%)
Median age (years)	39 (IQR: 20; range: 22-62)	40 (IQR: 18; range: 24-62)	36.5 (IQR: 22; range: 22-59)
≤30 – n (%)	24 (31%)	11 (30%)	13 (33%)
31-40 – n (%)	18 (23%)	9 (24%)	9 (23%)
41-50 – n (%)	18 (23%)	11 (30%)	7 (18%)
≥51– n (%)	17 (22%)	6 (16%)	11 (28%)
Median experience in hematology (years)	7.5 (IQR: 19; range: 0-34)[$]	7 (IQR: 14; range: 0-34)	8 (IQR: 21; range: 0-32)[$]
Median experience in HCT (years)	6 (IQR: 11; range: 0-32)[$]	6 (IQR: 12; range: 0-32)	6 (IQR: 11; range: 0-32)[$]
Median number of HCT patients evaluated for GvHD per week	1 (IQR: 5; range: 0-30)[$$]	1 (IQR: 5; range: 0-30)	1 (IQR: 5; range: 0-25)[$$]
very low (<1) - n (%)	25 (33%)	13 (35%)	12 (32%)
low (1-6) - n (%)	38 (51%)	17 (46%)	21 (55%)
moderate (7-15) - n (%)	6 (8%)	4 (11%)	2 (5%)
high (>15) - n (%)	6 (8%)	3 (8%)	3 (8%)
Area of expertise			
Adults only - n (%)	67 (87%)	32 (86%)	35 adults only (87%)
Children only – n (%)	2 (2%)	2 (5%)	0 children only (0%)
Both adults and children – n (%)	7 (9%)[$]	3 (8%)	4 both (10%)[$]
Median proficiency in English (1-10)	7 (IQR: 1; range: 2-10)[$$]	7.5 (IQR: 2; range: 2-10)[$]	7 (IQR: 1; range: 3-10)[$]

LEGEND

CI: confidence interval; IQR: inter quartile range; HCT: hematopoietic stem cell transplantation; OR: Odds ratio; $ = one respondent missing; * two nurses; ** one nurse and one medical student

Table 2: Survey 1 Results : Pre-test Practice Patterns, Technology Access and Technology Acceptance Data

Practice Patterns in GvHD assessment	Whole group (n= 77)	APP (n=37)	No APP (n=40)
Most often used International Guidelines* - n (%)			
Glucksberg criteria	**24 (31%)**	**12 (32%)**	**12 (30%)**
IBMTR Criteria	5 (7%)	2 (5%)	3 (8%)
MAGIC criteria	13 (17%)	4 (11%)	9 (23%)
Seattle Criteria	13 (17%)	6 (16%)	7 (18%)
NIH 2005 Criteria	14 (18%)	5 (14%)	9 (23%)
NIH 2014 Criteria	**27 (35%)**	**17 (46%)**	**10 (26%)**
Other / Does not know	11 (14%)	7 (19%)	4 (10%)
Reported level of ... (Likert scale 1 (lowest)-10 (highest))			
Median **importance** of the guidelines	7 (IQR 4 - range: 1-10)$^{\$\$\$\$\$\$}$	6 (IQR 4 - range: 1-10)$^{\$\$\$}$	7 (IQR 5 - range: 1-10)$^{\$\$\$}$
Median **comfort** in applying the guidelines	5 (IQR 3 - range: 1-9)$^{\$\$\$}$	5 (IQR 4 - range: 1-9)$^{\$\$}$	5 (IQR 3 - range: 1-9)$^{\$}$
Low (≤ 4) – n (%)	31 (42%)	17 (49%)	14 (35%)
Moderate (5-7) – n (%)	35 (47%)	14 (40%)	21 (54%)
High (≥ 8) – n (%)	8 (11%)	4 (11%)	4 (10%)
In my daily practice, my GvHD assessment relies on…* - n (%)			
Own knowledge	**44 (57%)**	**18 (50%)**	**26 (65%)**
A self-designed paper form	16 (21%)	7 (19%)	9 (23%)
A self-designed electronic file	5 (7%)	2 (5%)	3 (8%)

The official NIH 2005 paper form	8 (10%)	3 (8%)	5 (13%)
The official NIH 2014 paper form	17 (22%)	10 (27%)	7 (18%)
Other	14 (18%)	8 (22%)	6 (15%)
Not answered	2 (3%)	1 (3%)	1 (3%)
During the study, my GvHD assessment will rely on…* - n (%)			
Own knowledge	**NA**	**NA**	**24 (62%)**
A self-designed paper form	NA	NA	6 (15%)
The official NIH 2005 paper form	NA	NA	6 (15%)
The official NIH 2014 paper form	NA	NA	9 (23%)
Other	NA	NA	7 (18%)
Not answered	NA	NA	1 (3%)
Technology Access			
To support my daily practice, I have access to…* - n (%)			
A desktop computer with no internet connection	7 (9%)	3 (8%)	4 (10%)
A desktop computer with an internet connection	**70 (91%)**	**34 (92%)**	**36 (90%)**
A portable device	33 (43%)	17 (46%)	16 (40%)
A WIFI connection	31 (40%)	13 (35%)	18 (45%)
An electronic patient medical file	**48 (62%)**	**24 (65%)**	**24 (60%)**

Other	2 (3%)	0 (0%)	2 (5%)
Not answered	3 (4%)	1 (3%)	2 (5%)
Predicted use			
Localization of use* - n (%)			
Bedside	23 (30%)	12 (32%)	11 (28%)
Deskside	**57 (74%)**	**27 (73%)**	**30 (75%)**
Unlikely to use	2 (3%)	2 (5%)	0 (0%)
Other	1 (1%)	0	1 (3%)
Not answered	2 (3%)	0	2 (5%)
Type of device used* - n (%)			
Cellphone	**43 (56%)**	**25 (68%)**	**18 (45%)**
Tablet	5 (7%)	2 (5%)	3 (8%)
Laptop	6 (8%)	3 (8%)	3 (8%)
Desktop	**32 (42%)**	**10 (27%)**	**22 (55%)**
Other	0 (0%)	1 (0%)	0 (0%)
Not answered	2 (3%)	0	2 (5%)
Language			
Median importance of the availability of the app in my **native language** (Likert scale 1 (lowest)-10 (highest))	4 (IQR 5; range: 1-10)[$$]	4 (IQR 6; range: 1-10)	4 (IQR 5; range: 1-10)[$$]
Technology Acceptance Data			
Median reported level of **likelihood of using** the app (Likert scale 1 (lowest)-10 (highest))	8 (IQR 3; range: 1-10)[$$$$$$]	7.5 (IQR 3; range: 1-10)[$]	8 (IQR 4; range: 1-10)[$$$$]

LEGEND

CI: confidence interval; IQR: inter quartile range; NA: not applicable; OR: Odds ratio; * several answers were possible; $ = one participant missing

Table 3: GvHD Assessment Package Accuracy and Timing Results

Results for the complete GvHD test package (median)	APP (n=37)	No APP (n=40)
Correctly Diagnosed Vignettes *(maximum 10 correct answers)*	10 (IQR 1; range 5-10)	6.5 (IQR 3; range 2-9)
Correctly Scored Vignettes *(maximum 10 correct answers)*	9 (IQR 2; range 2-10)	4.5 (IQR 3; range 1-7)
Results for acute and chronic GvHD (median)	**APP (n=37)**	**No APP (n=40)**
Correctly Scored acute GvHD Vignettes *(maximum 4 correct answers)*	4 (IQR 0; range 2-4)	2 (IQR 2; range 0-4)
Correctly Scored chronic GvHD Vignettes *(maximum 6 correct answers)*	5 (IQR 1; range 0-6)	3 (IQR 2.25; range 0-5)
Time needed to complete the whole GvHD test package	**APP (n=37)**	**No APP (n=40)**
Mean time to complete all Vignettes (minutes)	48.84 (Std dev: 10.3; range 31-67)	25.27 (Std dev: 9.76; range 9-54)

LEGEND

CI: confidence interval; IQR: inter quartile range; Std dev: standard deviation

Table 4: Detailed Results of Participants for GvHD vignettes Compared to the Expert Gold Standard

	Results from the "App" group given by 37 participants - n (%)						
Expert Gold Standard Diagnosis	Classic Acute	Late Acute	Classic Chronic	Overlap Chronic	Missing	Other	Total
Classic acute GVHD °°	67 (91%)	4 (5%)	0 (0%)	2 (3%)	1 (1%)	0 (0%)	74 (20%)
Late acute GVHD °°	5 (7%)	65 (88%)	1 (1%)	3 (4%)	0 (0%)	0 (0%)	74 (20%)
Classic Chronic GVHD °°°°	3 (2%)	0 (0%)	140 (95%)	3 (2%)	2 (1%)	0 (0%)	148 (40%)
Overlap Chronic GVHD °°	0 (0%)	0 (0%)	4 (5%)	69 (93%)	0 (0%)	1 (1%)	74 (20%)
Total	75 (20%)	69 (18%)	145 (39%)	77 (21%)	3 (1%)	1 (0%)	370 (100%)
	Results from the "No App" group given by 40 participants - n (%)						
Expert Gold Standard Diagnosis	Classic Acute	Late Acute	Classic Chronic	Overlap Chronic	Missing	Other	Total
Classic acute GVHD °°	76 (95%)	0 (0%)	1 (1%)	2 (3%)	1 (1%)	0 (0%)	80 (20%)
Late acute GVHD °°	7 (9%)	52 (65%)	16 (20%)	5 (6%)	0 (0%)	0 (0%)	80 (20%)
Classic Chronic GVHD °°°°	18 (11%)	9 (6%)	110 (69%)	23 (14%)	0 (0%)	0 (0%)	160 (20%)
Overlap Chronic GVHD °°	3 (4%)	10 (13%)	51 (64%)	16 (20%)	0 (0%)	0 (0%)	80 (20%)
Total	104 (26%)	71 (18%)	178 (44%)	46 (11%)	1 (0%)	0 (0%)	400 (100%)

	Results from the "App" group given by 37 participants - n (%)									
Expert Gold Standard Severity Scoring	Grade I	Grade II	Grade III	Grade IV	Mild	Moderate	Severe	Missing	Other	Total
Grade I °	33 (89%)	1 (3%)	0 (0%)	0 (0%)	2 (5%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	37 (10%)
Grade II °	0 (0%)	37 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	37 (10%)
Grade III °	0 (0%)	0 (0%)	35 (95%)	0 (0%)	0 (0%)	0 (0%)	2 (5%)	0 (0%)	0 (0%)	37 (10%)
Grade IV °	0 (0%)	1 (3%)	3 (8%)	33 (89%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	37 (10%)
Mild °°	0 (0%)	0 (0%)	0 (0%)	0 (0%)	49 (66%)	22 (30%)	1 (1%)	1 (1%)	1 (1%)	74 (20%)
Moderate °°	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	61 (82%)	11 (15%)	1 (1%)	1 (1%)	74 (20%)
Severe °°	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (3%)	18 (24%)	54 (73%)	0 (0%)	0 (0%)	74 (20%)
Total	33 (9%)	39 (10%)	38 (10%)	33 (9%)	53 (14%)	102 (27%)	68 (18%)	2 (0%)	2 (0%)	370 (100%)
	Results from the "No App" group given by 40 participants - n (%)									
Expert Gold Standard Severity Scoring	Grade I	Grade II	Grade III	Grade IV	Mild	Moderate	Severe	Missing	Other	Total
Grade I °	29 (73%)	4 (10%)	0 (0%)	0 (0%)	4 (10%)	1 (3%)	0 (0%)	0 (0%)	2 (5%)	40 (10%)
Grade II °	3 (8%)	11 (28%)	4 (10%)	1 (3%)	3 (8%)	13 (33%)	4 (10%)	1 (3%)	0 (0%)	40 (10%)
Grade III °	0 (0%)	0 (0%)	27 (68%)	9 (23%)	0 (0%)	1 (3%)	1 (3%)	0 (0%)	2 (5%)	40 (10%)
Grade IV °	1 (3%)	8 (20%)	19 (48%)	7 (28%)	1 (3%)	2 (5%)	0 (0%)	0 (0%)	2 (5%)	40 (10%)
Mild °°	13 (16%)	12 (15%)	0 (0%)	0 (0%)	32 (40%)	19 (24%)	2 (1%)	1 (1%)	1 (1%)	80 (20%)
Moderate °°	5 (6%)	8 (10%)	4 (5%)	0 (0%)	5 (6%)	40 (50%)	15 (19%)	0 (0%)	3 (4%)	80 (20%)
Severe °°	1 (1%)	9 (11%)	9 (11%)	1 (1%)	8 (10%)	23 (29%)	27 (34%)	0 (0%)	2 (3%)	80 (20%)
Total	52 (13%)	52 (13%)	63 (16%)	18 (4.5%)	53 (13%)	99 (25%)	49 (12)	2 (0%)	12 (3%)	400 (100%)

LEGEND

NA = Not Applicable; vs. = versus; ° one clinical vignette; "Missing" corresponds to a lack of answer; "Other" corresponds to any answer not matching the proposed choices; The highlighted diagonal corresponds to the perfect agreement between participants and expert results.

Figures

Fig 1 – Study Design

Legend: APP: eGVHD App; GvHD: graft versus host disease

Fig 2 – CONSORT flow diagram

Legend: APP: eGVHD App; HCP: healthcare professional; HCT: hematopoietic stem cell transplantation

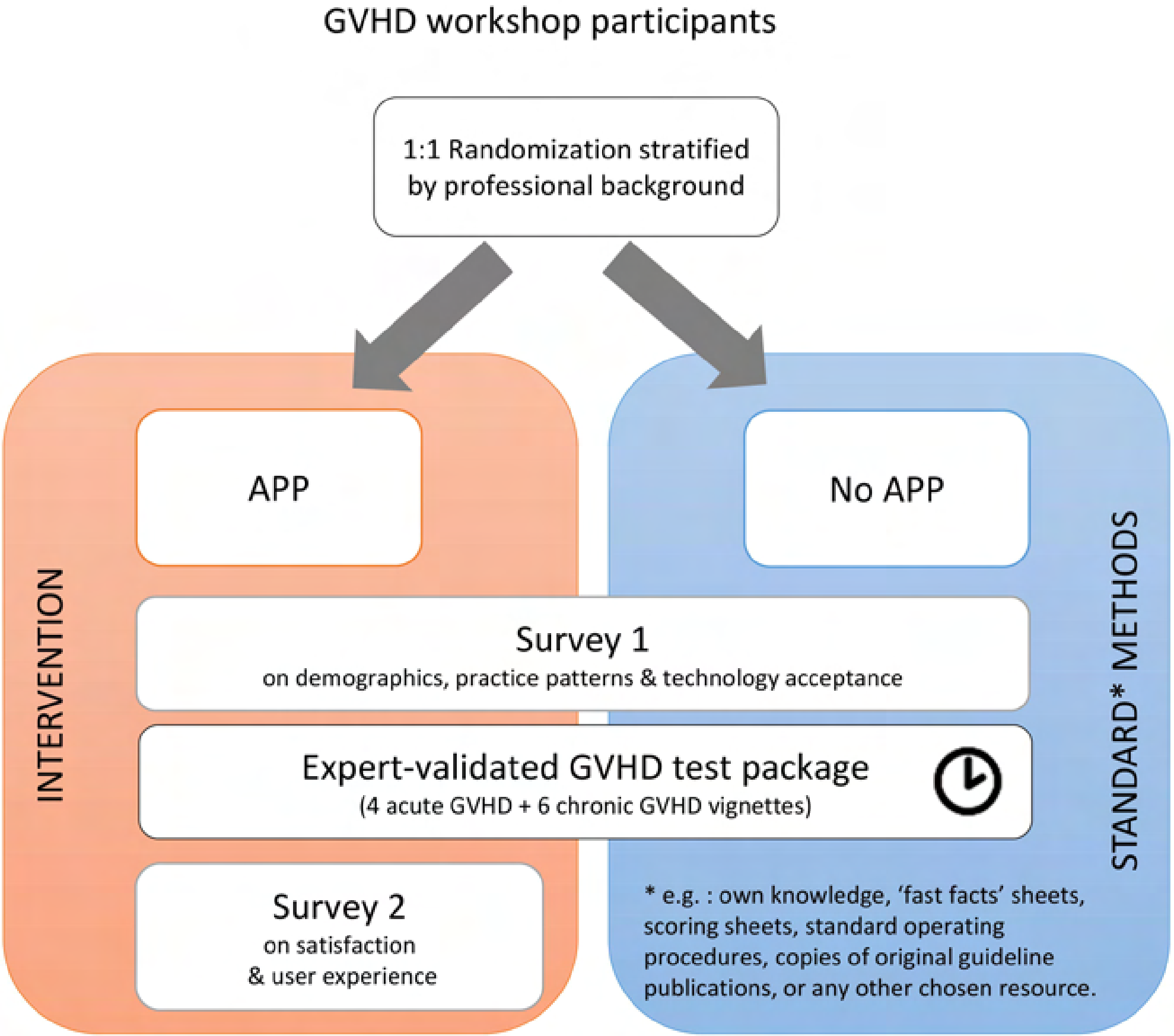
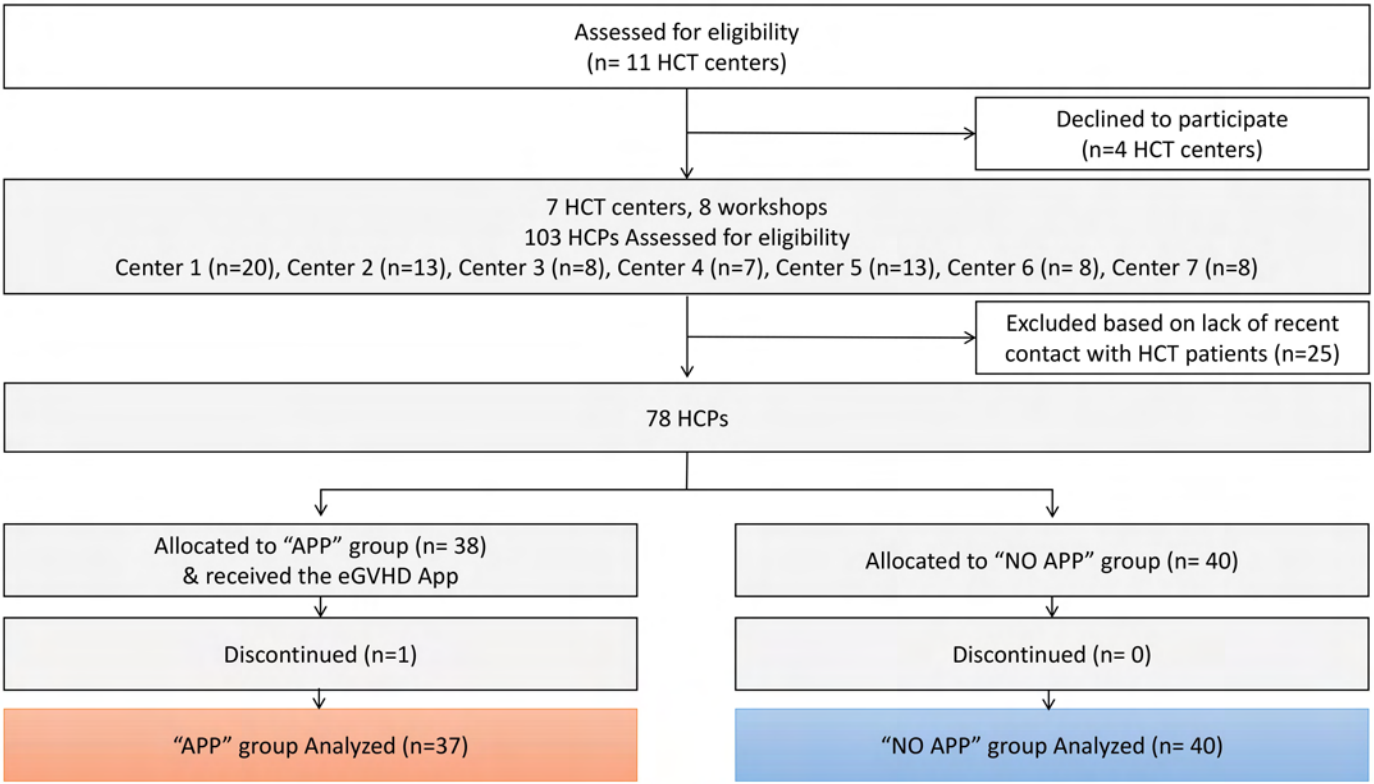

GVHD workshop participants
1:1 Randomization stratified by professional background
APP
No APP
INTERVENTION
STANDARD* METHODS
Survey 1
on demographics, practice patterns & technology acceptance
Expert-validated GVHD test package
(4 acute GVHD + 6 chronic GVHD vignettes)
Survey 2
on satisfaction
& user experience
* e.g. : own knowledge, 'fast facts' sheets, scoring sheets, standard operating procedures, copies of original guideline publications, or any other chosen resource.

Figure 2


Assessed for eligibility
(n= 11 HCT centers)
Declined to participate
(n=4 HCT centers)
7 HCT centers, 8 workshops
103 HCPs Assessed for eligibility
Center 1 (n=20), Center 2 (n=13), Center 3 (n=8), Center 4 (n=7), Center 5 (n=13), Center 6 (n= 8), Center 7 (n=8)
Excluded based on lack of recent contact with HCT patients (n=25)
78 HCPs
Allocated to "APP" group (n= 38)
& received the eGVHD App
Allocated to "NO APP" group (n= 40)
Discontinued (n=1)
Discontinued (n= 0)
"APP" group Analyzed (n=37)
"NO APP" group Analyzed (n= 40)

Supplemental methods

Sample and setting – Workshop description

The workshop lasted 90 to 120 minutes. It consisted of an introduction to the eGVHD App project, a short tutorial to train the "APP" group in the basic functionalities of the eGVHD App and the completion of a demographics, practice patterns and technology acceptance survey ("Survey 1") by all participants. Both groups then received the GvHD test package and recorded their answers individually. This was followed by a usability survey ("Survey 2") restricted to the "APP" group. The workshop concluded with a discussion of the correct answers of the test package and a summary of the most current recommendations for GvHD assessment.

Data collection points, randomization procedure and blinding

All data collection was performed during the workshop using pen and paper. Allocation to the intervention arm ("APP" group) was random and stratified. More specifically, randomization was done at the arrival of study participants based on pre-formatted randomization sheets (www.randomization.com) and order of arrival. We used randomly permuted blocks, with block sizes of 2, to compensate for the low number of participants per center. Stratification was based on professional background: (1) senior physicians (board certified hematologists), (2) junior physicians (medical doctors training in internal medicine or hematology), (3) data managers or research nurses specialized in HCT data entry or (4) other (e.g. medical students or nurses with no specific GvHD evaluation expertise). Blinding was not feasible due to the nature of the intervention.

Outcome measures – Planned sub-analyses

We planned the following sub-analyses: (1) to compare the difference in diagnosis accuracy between both groups, (2) to test for the App-effect on the accuracy of the severity scoring conditional on a GvHD diagnosis being acute (aGvHD) and chronic (cGvHD), (3) to verify whether the effect of the App depended on the type of GvHD, the severity of GvHD, professional background or center, (4) to compare the inter-rater reliability and (5) to compare the time needed to complete the full test package between both groups.

Variables and measurements

Gold Standard

Prior to this study, four GvHD experts (SZP, DW, AI and SJL) determined the correct diagnosis and severity score of each vignette, based on the MAGIC criteria[1] for acute GvHD and the NIH 2014 guidelines[2] for chronic GvHD by evaluating the ten clinical vignettes independently and returning their GvHD assessment separately to the principal investigator. The 'gold standard' for diagnosis and severity scoring corresponded to the answer given by at least three of the four experts. When an expert disagreed with the consensus of the other three experts, this expert was contacted separately to confirm that he/she agreed with the 'gold standard' answer given by the rest of the group.

Experts were healthcare professionals active in the field of allogeneic HCT, co-authors of at least one publication in the field of clinical GvHD and active members of an international GvHD consortium or working group.

Post-test user satisfaction and experience

Briefly, the TAM consists of six statements, referring to the extent to which the user believes the technology will improve his work performance. Statements are rated on a 7-point Likert-like scale (1= 'extremely unlikely' to 7= 'extremely likely'). A median score is calculated for each item separately, with higher scores reflecting higher perceived usefulness. The PSSUQ is a 19-item questionnaire using 7-point Likert-like scales (1= 'strongly agree' to 7= 'strongly disagree'), with three subscales reflecting system usefulness (items 1-8), information quality (items 9-15) and interface quality (items 16-18), respectively. PSSUQ scores are presented as median total and subscale scores, with lower scores reflecting higher user satisfaction.

Statistical analysis

Results were analyzed with IBM SPSS statistics version 24 and R version 3.3.3 according to the 'intention to treat' principle. Missing results were reported as such. Descriptive results were reported using a measure of central tendency and a measure of dispersion, as appropriate. The probability of a correct answer was compared between both groups using a mixed effects logistic regression model, for diagnosis and severity score separately. The model contained *fixed effects* of group ("App" versus "No App") and professional background (the stratification variable in the randomization) and *random effects* of center and workshop participant. These random effects were included to handle the correlation between the workshop

participants belonging to the same center, and between the ten answers given by the same workshop participant, respectively. Odds ratio's and 95% confidence intervals (CI) for the effect of group were reported. To verify whether the effect of the App depended on the type of the GvHD, the severity of the GvHD, professional background or center, interaction terms were added in separate models. *Inter-observer agreement* of the severity was evaluated by using the Brennan-Prediger's kappa coefficient (K_{BP}) which ranges between zero (no agreement) and one (perfect agreement). This coefficient evaluates the raters' agreement for nominal scales with more than two categories and takes into account the fact that agreement could have occurred by chance. This version of the kappa is reported instead of the classical Fleiss-Cohen kappa, since the latter is not appropriate for comparisons of conditions having a difference in distribution[3]. Kappa's are compared between both groups using an approach presented by Gwet and colleagues[4]. The *time needed to score* the vignettes was compared between both groups using a linear mixed model, with the same fixed and random effects as in the aforementioned logistic regression model.

Supplemental Tables

Supplementary Table 1:

Characteristics of Centers Performing Allogeneic HCT in Belgium and Participating in the Study

Center	Academic Center	total HCT per year	1st alloHCT per year	% activity in Belgium	total number of participants	Senior MD	Junior MD	Data managers	Other
1	yes	130	76	82%	20	8	8	4	0
2	yes	95	57		13	5	4	3	1
3	no	87	33		8	3	2	1	2
4	yes	74	34		7	4	1	1	1
5	yes	79	49		13	6	3	4	0
6	yes	57	20		8	7	0	1	0
7	yes	93	43		8	4	3	1	0
8	no	39	17	18%	Declined	NA	NA	NA	NA
9	yes	47	26		Declined	NA	NA	NA	NA
10	yes	31	12		Declined	NA	NA	NA	NA
11	no	35	15		Declined	NA	NA	NA	NA

LEGEND

HCT: hematopoietic stem cell transplantation; alloHCT: allogeneic hematopoietic stem cell transplantation; MD: medical doctor

Supplementary Table 2: List of Clinical Vignettes

Vignette	Decription of the Clinical Vignette	Diagnosis	Severity Scoring
1	1. A female adult patient receives an allogeneic stem cell transplantation for a myelodysplasia. Her post transplantation period is uneventful, but 9 months after transplantation she develops: • a red inflammatory rash on both arms, one month after discontinuation of immunosuppression. • There are no other abnormal signs or symptoms and her pulmonary function lab results are normal. • A biopsy of the skin of the forearm is suggestive for GVHD (likely GVHD - apoptosis in epidermal basal layer).	Late acute GVHD	Grade I
2	A female adult patient receives an allogeneic stem cell transplantation for a chronic myeloid leukemia. Her pre-transplantation evaluation is unremarkable. Around day 90, she develops: • dyspnea when walking on flat ground. • A pulmonary evaluation reveals a newly decreased FEV1* of 65%, with a FEV1/VC ratio** of 0.65 and a RV*** of 110%. • Air trapping is present on high resolution CT scan of the lungs. • Infections of the respiratory tract are excluded by a normal bronchial aspirate evaluation and her cardiac function is normal. • Her clinical exam and laboratory results are perfectly normal except for xerostomia (dry mouth), without impact on her oral intake.	Classic Chronic GVHD	Moderate

3	Four months after receiving an allogeneic stem cell transplantation, a female adult patient presents with: • anorexia, daily vomiting and an unintentional weight loss of about 15% of her pre-transplantation weight. • There are no other abnormal signs or symptoms and her lab results and pulmonary function tests are normal. • A stomach biopsy confirms GVHD (likely GVHD - gastric pit apoptosis).	Late acute GVHD	Grade II
4	Four weeks after receiving an allogeneic stem cell transplantation, a male adult patient presents with: • an itchy erythematous rash involving the head and neck, and anorexia with major diarrhea (10x/day, about 2000ml/day) but no abdominal pain. • A colonoscopy confirms GVHD by biopsy (likely GVHD - crypt apoptosis in the intestines) and excludes a concomitant infection or drug toxicity. • His lab results are normal except for a low albumin and slightly elevated creatinine. • His pulmonary function tests are normal.	Classic acute GVHD	Grade III
5	A female adult patient, 6 months after her allogeneic stem cell transplantation, develops: • two patches of morphea-like lesions (patches of leather-like, shiny skin) on the lower back (diameter 5cm) • with an elevation of liver enzymes (ALT, AST, AP and GGT a little more than 3x the upper normal limit), without other potential confounding cause. • Her pulmonary function tests are normal. • She reports dyspareunia (painful intercourse) and a gynecological exam reveals vaginal adhesions and scarring.	Overlap Chronic GVHD	SEVERE

6	Two months after receiving an allogeneic stem cell transplantation, a male adult patient develops: • relatively frequent diarrhea episodes (4 times/ a day) accompanied with very severely painful abdominal cramps. • A colonoscopy confirms GVHD by biopsy (likely GVHD – apoptosis in enterocytes and destruction of crypt architecture) and excludes a concomitant infection / drug toxicity. • His body weight is unchanged. • Liver enzymes are slightly elevated (ALT, AST, AP and GGT slightly more than twice the upper normal limit) and bilirubin is 3.5 mg/dL, without argument for infection, drug toxicity or veno-occlusive disease. • Except for fatigue, there are no other abnormal signs or symptoms and the rest of his lab results are normal. • His pulmonary function tests are normal.	Classic acute GVHD	Grade IV

7	A male adult patient receives an allogeneic stem cell transplantation. Ten months later, he feels fine but he reports: • frequent muscle cramps and has noticed that some movements have become more difficult : an increasing tightness in his lower back, arms and legs is making it more difficult to his daily jog and pick up items on the ground. • On clinical exam, the extension of the arms and the flexion of the wrist are somewhat decreased and the ankles moderately swollen without inflammatory features. • No other clinical abnormalities are found than new lichen sclerosus-like changes (white patches of firm thickened/crinkled skin with a tendency to scar) that have appeared on the penis. • An electromyography is normal. His laboratory exams are normal, including muscle enzymes. • His pulmonary function tests are normal.	Classic Chronic GVHD	Moderate
8	Three months after her transplantation, a female adult allogeneic transplantation recipient presents with: • two new painful ulcerations in the mouth. Oral exam reveals lichen planus like changes and 1.5cm wide ulcerations. Microbial examination for candida and herpes are negative. • Weight is stable but the patient no longer tolerates sparkling drinks. Oral intake is preserved. • The rest of her clinical exam, pulmonary function tests and lab results are unremarkable.	Classic Chronic GVHD	MILD
9	A male adult patient, five months after receiving an allogeneic stem cell transplantation, develops: • a new maculopapular inflammatory red rash on the hands and feet.	Overlap Chronic GVHD	MILD

	• He also notices that his nails have become brittle and his eyes are more sensitive than before. • The ophthalmologist confirms signs of keratoconjunctivitis sicca with a slit lamp examination. Shirmer's test shows a 3mm tear production after 5 minutes. • His ocular problems are totally relieved by using artificial teardrops twice a day. • Further exams reveal normal pulmonary function tests, an unremarkable clinical exam (except for the rash and dystrophic nails). • He has normal laboratory results except for slightly elevated alkaline phosphatase (AP) which are a little over twice the upper normal limit (2x ULN).		
10	Twelve months after her allogeneic stem cell transplantation, a female adult patient develops: • difficulties with swallowing due to non-painful xerostomia (dry mouth) and the impression that food remains stuck when she swallows. • Her weight remains stable but she needs to chew abnormally long and drink along almost all of her solid food intakes. • A gastroscopy confirms the presence of a new stenosis of the upper esophagus, which is successfully dilated but no biopsies are taken. • Her other clinical, laboratory and pulmonary function test evaluations are normal, except for some superficial sclerosis bilaterally in the lower arms and legs.	Classic Chronic GVHD	SEVERE

LEGEND

* FEV1 = forced expiratory volume in one second

**FEV1/VC =Tiffeneau index or forced expiratory volume in one second divided by vital capacity

***RV= residual volume

Supplementary Table 3: Full Binary (Correct vs Incorrect) Results of Participants for Individual GvHD vignettes

LEGEND

* missing results were considered as being incorrect; v.s.: versus; Δ: difference between

Supplementary Table 3: **Full Binary (correct vs. incorrect) Results of Participants for Individual GvHD Vignettes**

Vignette number	Gold Standard Diagnosis	Participant Diagnosis*	All participants n=77	"APP" group n=37	"NO APP" group n=40	Δ "App" vs. "No App" (%) Correct Diagnosis	Gold Standard Severity Scoring	Participant Severity Scoring	All participants n=77	"APP" group n=37	"NO APP" group n=40	Δ "App" vs. "No App" (%) Correct Severity Scoring
1	Late **acute** GVHD	correct	69 (89.6%)	33 (89.2%)	36 (90%)	- 0.8 %	Grade I	correct	62 (80.5%)	33 (89.2%)	29 (72.5%)	+ 16.7 %
		incorrect	8 (10.4%)	4 (10.8%)	4 (10%)			incorrect	15 (19.5%)	4 (10.8%)	11 (27.5%)	
3	Late **acute** GVHD	correct	60 (77.9%)	37 (100%)	23 (57.5%)	+ 42.5%	Grade II	correct	48 (62.3%)	37 (100%)	11 (27.5%)	+ 72.5%
		incorrect	17 (22.1%)	0 (0%)	17 (42.5%)			incorrect	29 (37.7%)	0 (0%)	29 (72.5%)	
4	Classic **acute** GVHD	correct	72 (93.5%)	34 (91.9%)	38 (95%)	- 3.1%	Grade III	correct	62 (80.5%)	35 (94.6%)	27 (67.5%)	+ 27.1%
		incorrect	5 (6.5%)	3 (8.1%)	2 (5%)			incorrect	15 (19.5%)	2 (5.4%)	13 (32.5%)	
6	Classic **acute** GVHD	correct	75 (97.4%)	37 (100%)	38 (95%)	+ 5%	Grade IV	correct	40 (51.9%)	33 (89.2%)	7 (17.5%)	+ 71.7%
		incorrect	2 (2.6%)	0 (0%)	2 (5%)			incorrect	37 (48.1%)	4 (10.8%)	33 (82.5%)	
8	Classic **Chronic** GVHD	correct	64 (83.1%)	35 (94.6%)	29 (72.5%)	+ 22.1%	Mild	correct	39 (50.6%)	19 (51.3%)	20 (50%)	+ 1.3%
		incorrect	13 (16.9%)	2 (5.4%)	11 (27.5%)			incorrect	38 (49.4%)	18 (48.6%)	20 (50%)	
9	Overlap **Chronic** GVHD	correct	68 (88.3%)	36 (97.3%)	32 (80%)	+ 17.3%	Mild	correct	42 (54.5%)	30 (81.1%)	12 (30%)	+ 51.1%
		incorrect	9 (11.7%)	1(2.7%)	8 (20%)			incorrect	35 (45.5%)	7 (18.9%)	28 (70%)	
2	Classic **Chronic** GVHD	correct	66 (85.7%)	35 (94.6%)	31 (77.5%)	+ 17.1%	Moderate	correct	48 (62.3%)	32 (86.5%)	16 (40%)	+ 46.5%
		incorrect	11 (14.3%)	2 (5.4%)	9 (22.5%)			incorrect	29 (37.7%)	5 (13.5%)	24 (60%)	
7	Classic **Chronic** GVHD	correct	72 (93.5%)	36 (97.3%)	36 (90%)	+ 7.3%	Moderate	correct	53 (68.8%)	29 (78.4%)	24 (60%)	+ 18.4%
		incorrect	5 (6.5%)	1 (2.7%)	4 (10%)			incorrect	24 (31.2%)	8 (21.6%)	16 (40%)	
5	Overlap **Chronic** GVHD	correct	72 (93.5%)	37 (100%)	35 (87.5%)	+ 12.5%	Severe	correct	41 (53.2%)	29 (78.4%)	12 (30%)	+ 48.4%
		incorrect	5 (6.5%)	0 (0%)	5(12.5%)			incorrect	36 (46.7%)	8 (21.6%)	28 (70%)	
10	Classic **Chronic** GVHD	correct	74 (96.1%)	37 (100%)	37 (92.5%)	+ 7.5%	Severe	correct	40 (51.9%)	25 (67.6%)	15 (37.5%)	+ 30.1%
		incorrect	3 (3.9%)	0 (0%)	3 (7.5%)			incorrect	37 (48.1%)	12 (32.4%)	25 (62.5%)	

LEGEND

* missing results were considered as being incorrect

v.s.: versus; Δ: difference between

Supplementary Table 4: Post-test User experience and Usability Data ("APP" group only)

LEGEND

IQR: inter quartile range; TAM: Technology Assessment Model; PSSUQ: Post-Study System Usability Questionnaire

Supplementary Table 4: Post-test User Experience and Usability Data ("APP" group only)

Perceived Usefulness – TAM median score (**7= extremely likely**; 1= extremely unlikely)		
Using the "EBMT GVHD app" would…	n	
Enable me to accomplish tasks more quickly.	37	5 (IQR 2; range: 2-7)
Improve my job performance.	37	6 (IQR 1; range: 5-7)
Increase my productivity.	37	5 (IQR 1; range: 3-7)
Enhance my effectiveness on the job.	37	6 (IQR 1; range: 4-7)
Make it easier to do my job.	37	6 (IQR 1; range: 3-7)
I would find the "EBMT GVHD app" useful in my job.	37	6 (IQR 1; range: 4-7)
System Usability – PSSUQ median score (**1= Strongly agree** ; 7= Strongly disagree)	n	
1. Overall, I am satisfied with how easy it is to use this system.	36	2 (IQR 1; range 1-3)
2. It was simple to use this system.	36	2 (IQR 1; range 1-5)
3. I could effectively complete the tasks and scenarios using this system.	36	2 (IQR 1; range 1-4)
4. I was able to complete the tasks and scenarios quickly using this system	36	3 (IQR 1; range 1-6)
5. I was able to efficiently complete the tasks and scenarios quickly using this system.	36	2 (IQR 1; range 1-6)
6. I felt comfortable using this system.	36	2 (IQR 2; range 1-5)
7. It was easy to learn to use this system.	36	1.5 (IQR 1; range 1-5)
8. I believe I could become productive quickly using this system.	36	2 (IQR 1; range 1-5)
System use subscale score		2 (IQR 1; range 1-5)
9. The system gave error messages that clearly told me how to fix problems.	26	2.5 (IQR2; range 1-6)
10. Whenever I made a mistake using the system, I could recover easily and quickly.	34	2 (IQR 1; range 1-5)
11. The information provided with this system was clear.	36	2 (IQR 0; range 1-3)
12. It was easy to find the information I needed.	29	2 (IQR 0; range 1-5)
13. The information provided for the system was easy to understand.	35	2 (IQR: 0; range 1-2)
14. The information was effective in helping me complete the tasks and scenarios.	35	2 (IQR: 0; range 1-3)
15. The organization of information on the system screens was clear.	36	2 (IQR: 1; range 1-6)
Information quality subscale score		2 (IQR 1; range 1-3)
16. The interface (= items you use to interact with the system e.g. screen, mouse, keyboard,…) of this system was pleasant.	35	2 (IQR: 1; range 1-6)
17. I liked using the interface of this system.	35	2 (IQR: 1; range 1-5)
18. This system has all the functions and capabilities I expect it to have.	36	2 (IQR: 1; range 1-3)
Interface quality subscale score		2 (IQR 1; range 1-3)
19. Overall, I am satisfied with this system.	36	2 (IQR: 1; range 1-3)
Overall PSSUQ (items 1-19)		2 (IQR 0,4; range 1-3)
Predicted use		
Reported level of **likelihood of using the app** in the future (Likert scale 1 (lowest)-10 (highest))	36	8 (IQR 3; range: 1-10)
Actual use		
Reported level of **comfort using the app** in English (Likert scale 1 (lowest)-10 (highest))	37	9 (IQR 2.5; range: 3-10)

LEGEND
IQR: inter quartile range; TAM: Technology Assessment Model; PSSUQ: Post-Study System Usability Questionnaire

Supplemental Figures

Supplementary Figure 1 – The superiority of the eGVHD App in GvHD assessment is similar regardless of **center effect** for diagnosis (A) and severity scoring (B)

Supplementary Figure 2 – The superiority of the eGVHD App in GvHD assessment is similar regardless of **professional background** for diagnosis (A) and severity scoring (B)

Supplementary Figure 3 – The superiority of the eGVHD App in GvHD assessment is similar regardless of the **age category** of the user for diagnosis (A) and severity scoring (B)

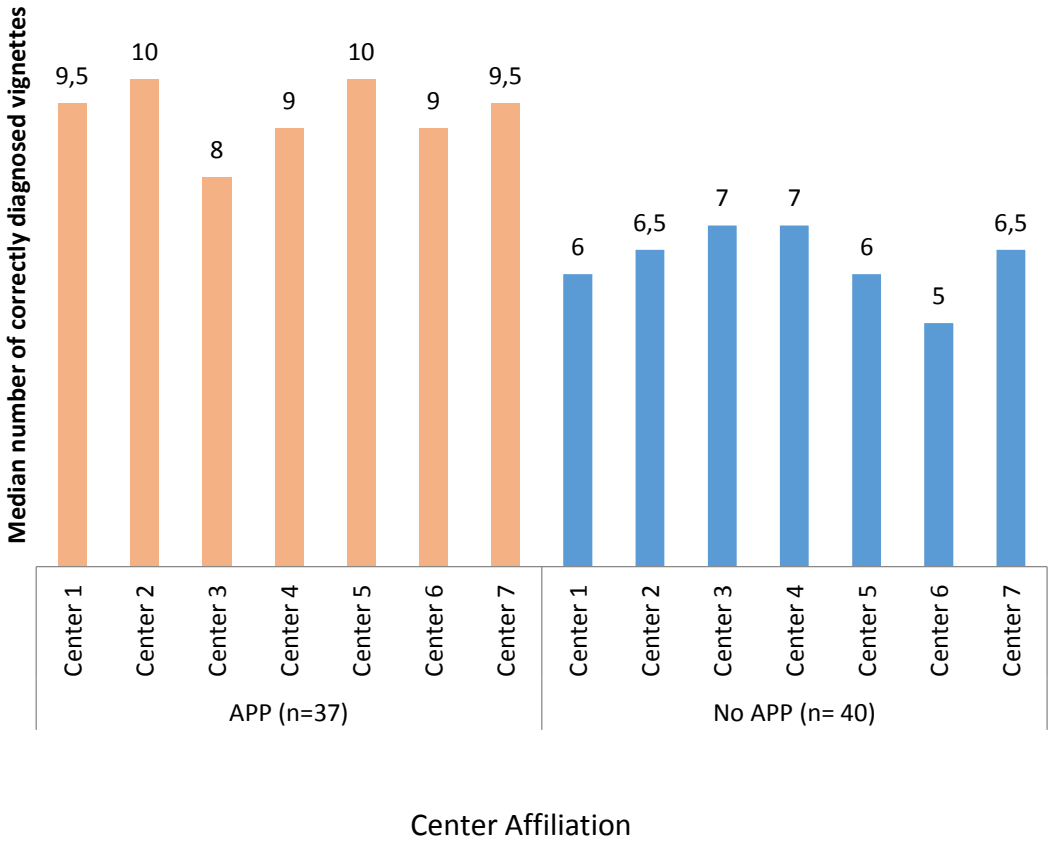
Supplementary Figure 4 – The superiority of the eGVHD App in GvHD assessment is similar regardless of the **user self-reported experience with GvHD assessment** for diagnosis (A) and severity scoring (B)

Legend supplementary Figure 4:
User experience with GvHD assessment was categorized according to the number of HCT patients the health care professionals reported to evaluate per week: "very low" experience (less than one HCT patient per week), "low" (1 to 6 weekly contacts with HCT patients), "moderate" experience (7 to 15 weekly contacts with HCT patients); "high" experience (more than 15 weekly contacts with HCT patients)

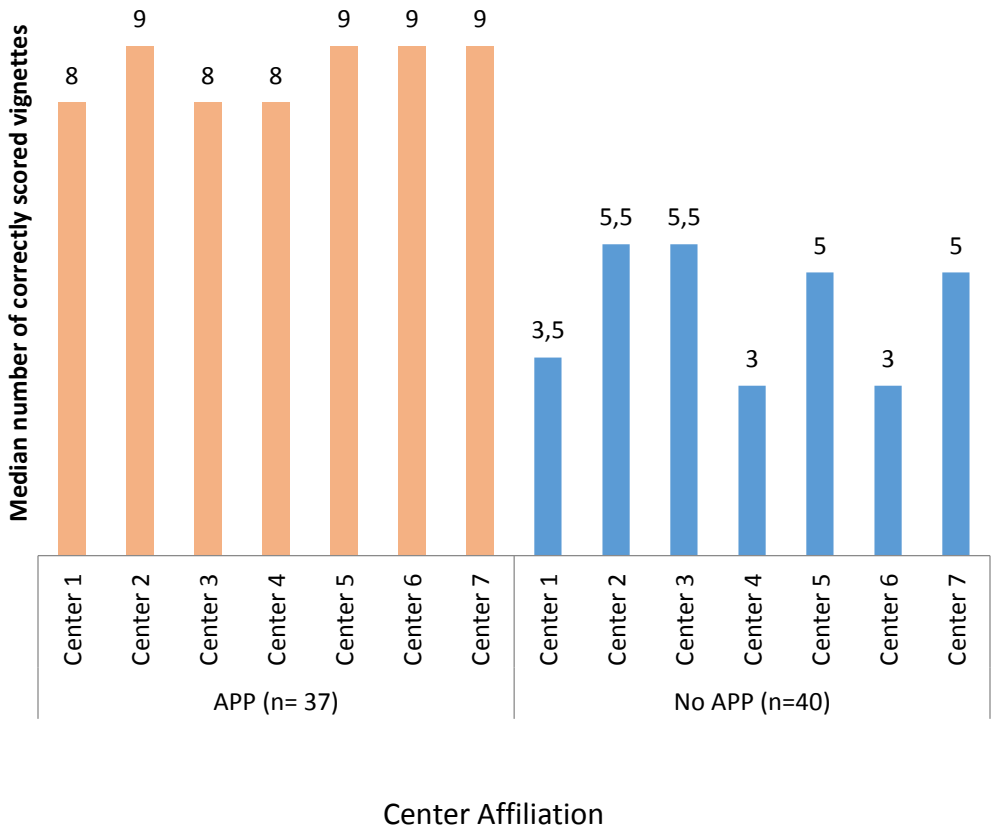
Supplementary Figure 5 – The superiority of the eGVHD App in GvHD assessment is similar regardless of the **user self-reported comfort with GvHD assessment guidelines** for diagnosis (A) and severity scoring (B)

Legend supplementary Figure 5:
User self-reported comfort with GvHD assessment guidelines was categorized based on the pre-test survey question "How comfortable are you with using the above mentioned criteria in your daily practice on a Likert scale of 1-10 (1= Not at all comfortable; 10= extremely comfortable)": "low" comfort (response 4 or less), "moderate" comfort (response between 5 and 7), "high" comfort (response 8 or above).

A. GvHD Diagnosis

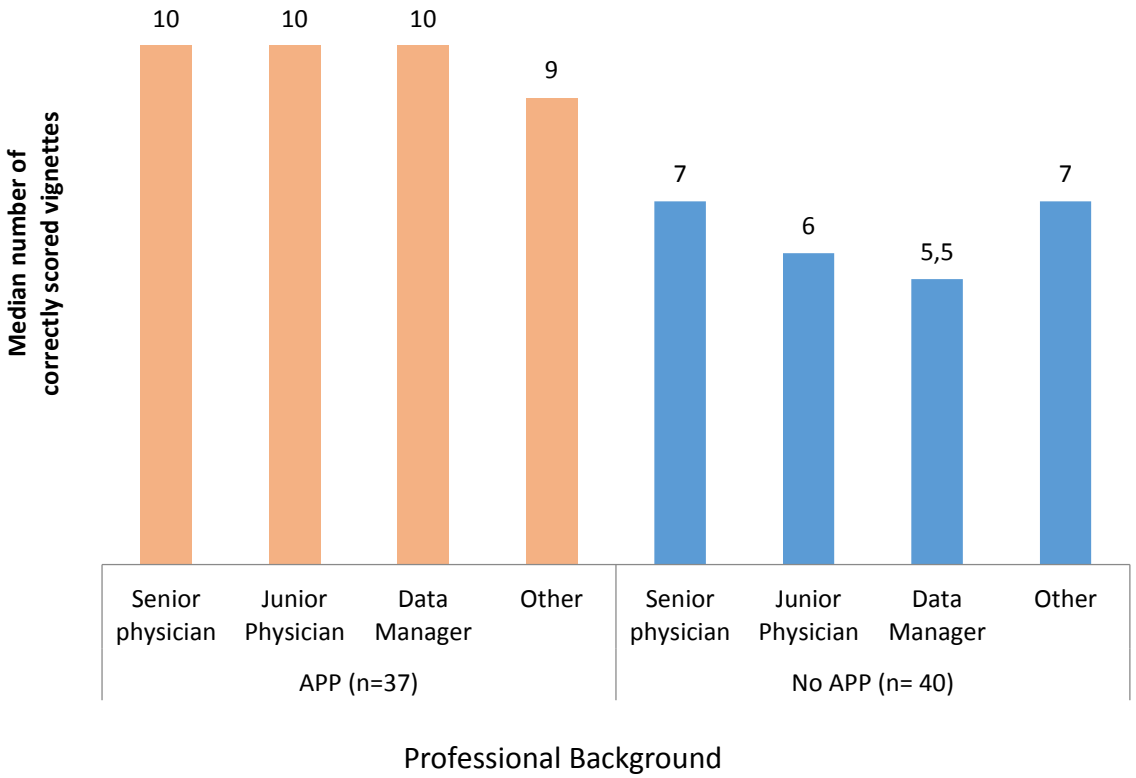


B. GvHD Severity Scoring



p=0.445

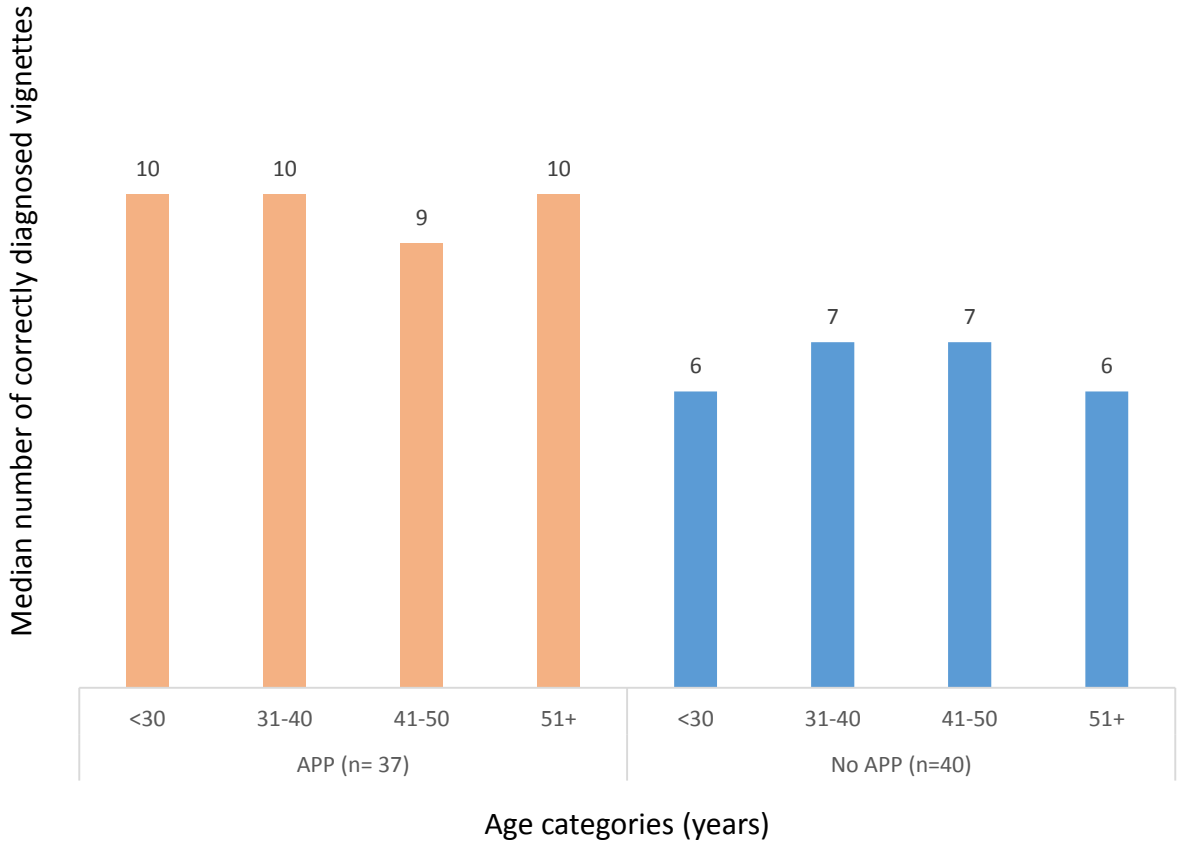
A. GvHD Diagnosis



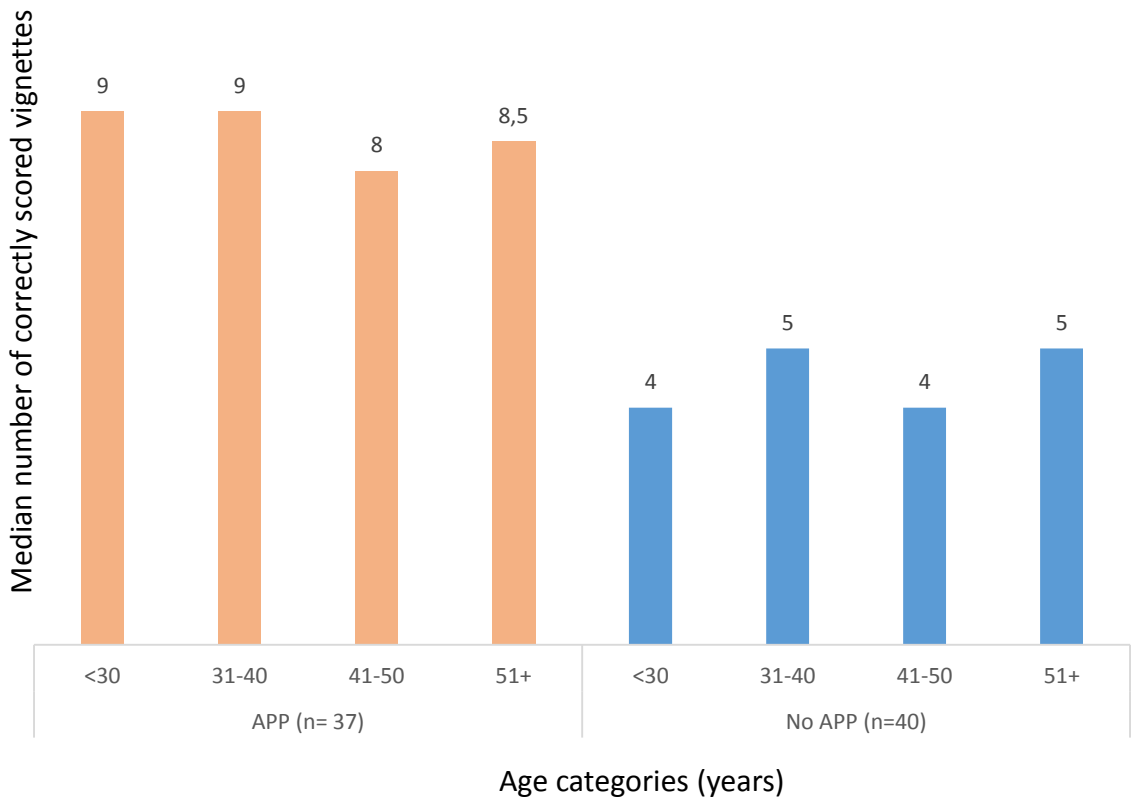
B. GvHD Severity Scoring

p=0.665

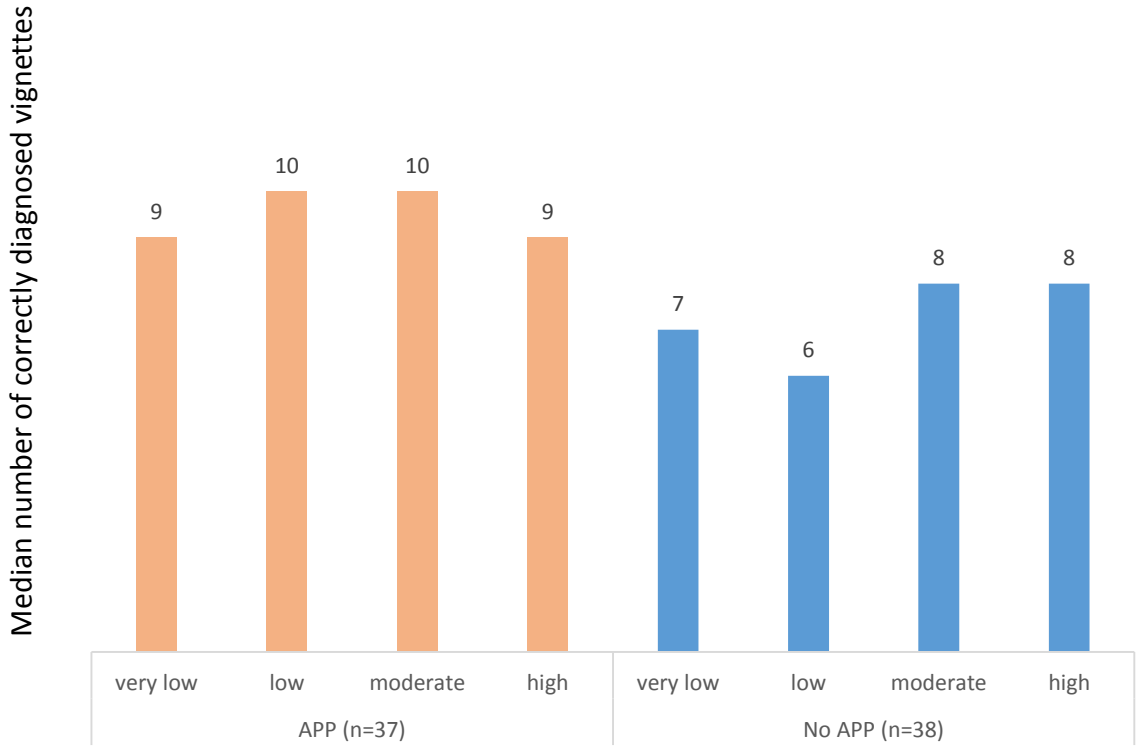
A. GvHD Diagnosis



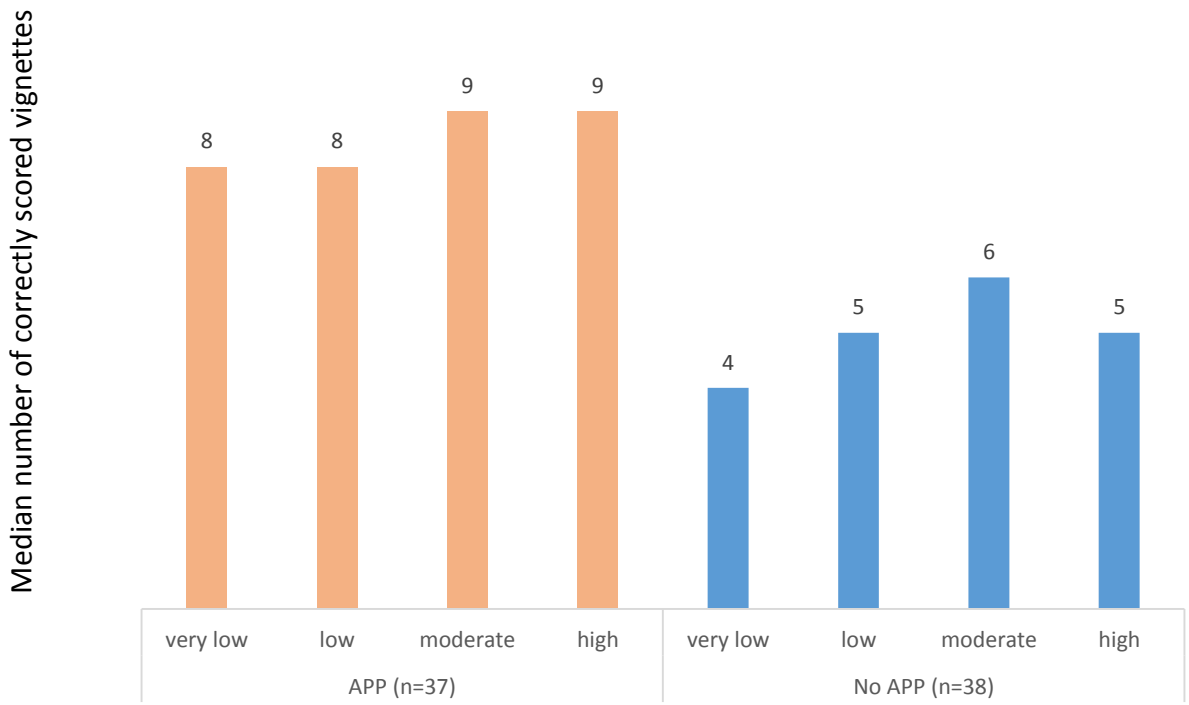
B. GvHD Severity Scoring



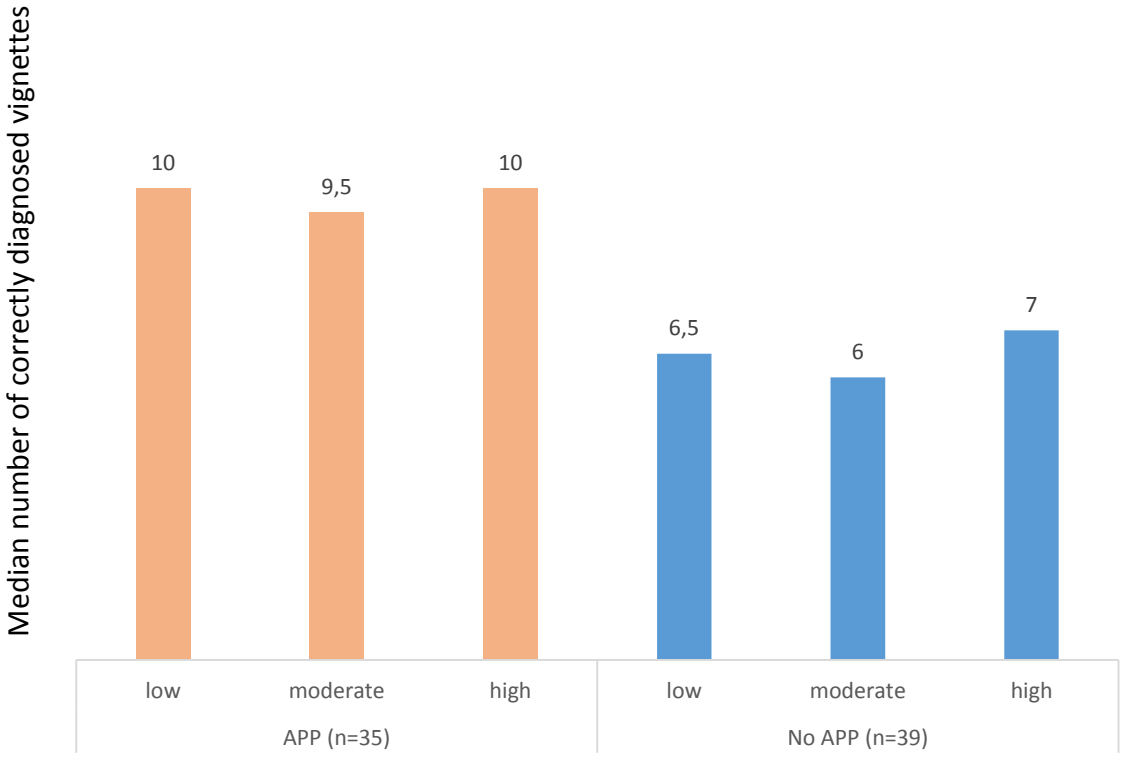
A. GvHD Diagnosis



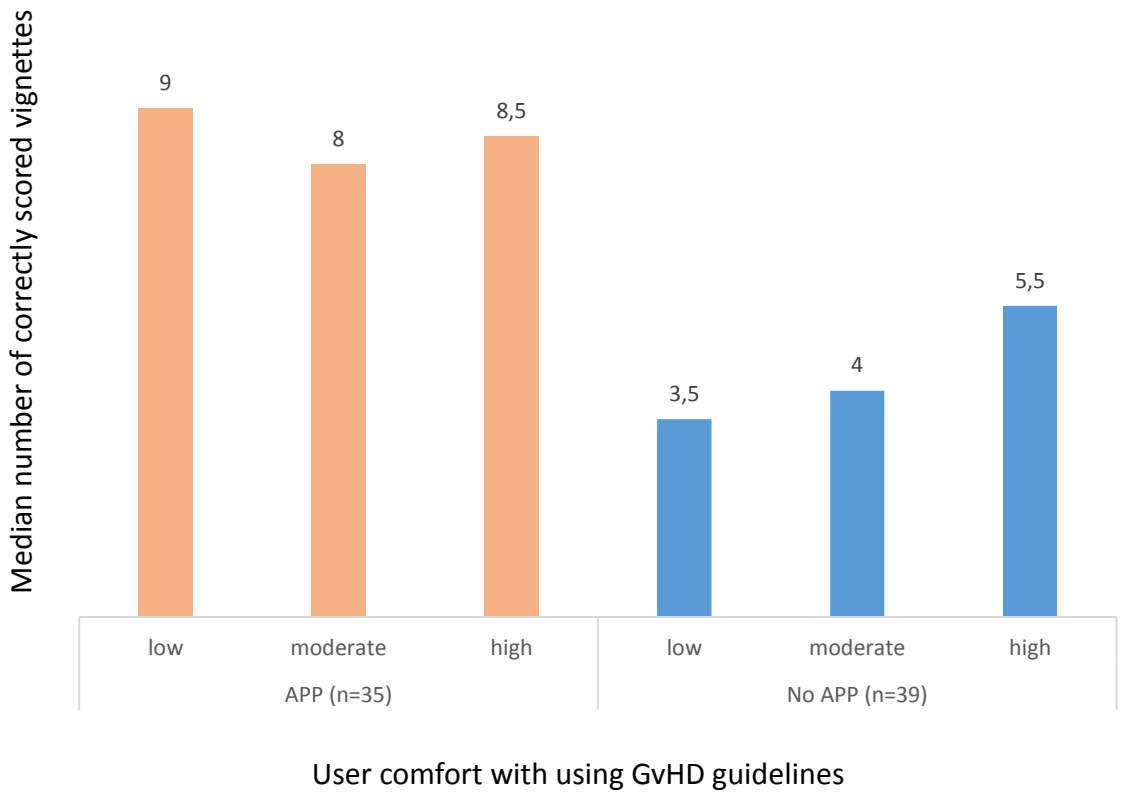
B. GvHD Severity Scoring



A. GvHD Diagnosis



B. GvHD Severity Scoring



Supplemental References

1. Harris AC, Young R, Devine S, et al. International, Multicenter Standardization of Acute Graft-versus-Host Disease Clinical Data Collection: A Report from the Mount Sinai Acute GVHD International Consortium. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2016;22(1):4-10.
2. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease: I. The 2014 Diagnosis and Staging Working Group report. Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation. 2015;21(3):389-401 e381.
3. von Eye A, von Eye M. On the marginal dependency of cohen's κ. European Psychologist 2008;13(305-315).
4. Gwet KL. Testing the difference of correlated agreement coefficients for statistical significance. Educational and Psychological Measurement 2016. 2016;76(609-637).