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,[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;[2] on behalf of the EBMT Transplantation Complications Working party

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

Haematologica 2018
Volume 103(10):1698-1707

Correspondence:
helene.schoemans@uzleuven.be

Received: March 6, 2018.
Accepted: June 13, 2018.
Pre-published: June 14, 2018.

doi:10.3324/haematol.2018.190777

Check the online version for the most updated information on this article, online supplements, and information on authorship & disclosures: www.haematologica.org/content/103/10/1698

©2018 Ferrata Storti Foundation

Material published in Haematologica is covered by copyright. All rights are reserved to the Ferrata Storti Foundation. Use of published material is allowed under the following terms and conditions: https://creativecommons.org/licenses/by-nc/4.0/legalcode. Copies of published material are allowed for personal or internal use. Sharing published material for non-commercial purposes is subject to the following conditions: https://creativecommons.org/licenses/by-nc/4.0/legalcode, sect. 3. Reproducing and sharing published material for commercial purposes is not allowed without permission in writing from the publisher.

ABSTRACT

Graft-*versus*-host disease (GvHD) 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 GvHD. Our national multicenter randomized controlled trial involved seven Belgian transplantation centers and 78 healthcare professionals selected using a 2-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 GvHD assessment aids ("No APP") to assess the diagnosis and severity score of 10 expert-validated clinical vignettes. Our main outcome measure was the difference in accuracy for GvHD 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 GvHD severity scoring was significantly superior for both acute and chronic GvHD, 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 the time needed to score the whole GvHD test package in the "APP" group (P<0.001), usability feedback was positive. The eGVHD App shows superior GvHD assessment accuracy compared to standard practice and has the potential to improve the quality of outcome data registration in allogeneic stem cell transplantation.

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 problems implementing GvHD assessment correctly, 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 European Group for Blood and Marrow Transplantation (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 the 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 the tool's usability to be further improved. 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 2-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 (*Online Supplementary Table S1*) 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 (from 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 *Online Supplementary Methods* for workshop details) and could recall having performed at least one GvHD evaluation in the past 12 months.

Information concerning data collection points, randomization procedure and blinding are available in the *Online Supplementary 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 *Online Supplementary 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 and acceptance data (Table 2) at baseline.

Accuracy of GvHD assessment: participants were required to diagnose and score a package of 10 randomly ordered GvHD clinical vignettes based on real-life clinical cases (see *Online Supplementary Methods* and *Online Supplementary Table S2*) according to the most up-to-date international guidelines.[1] Four acute GvHD (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 Mount Sinai Acute GvHD International Consortium (MAGIC) criteria.[8] Six chronic GvHD (cGvHD) vignettes covered the two cGvHD diagnoses ('overlap cGvHD' and 'classic cGvHD', two and four vignettes, respectively) and the three severity grades of the National Institutes of Health (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 *Online Supplementary 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 (*Online Supplementary Table S3*). 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 meth-

Table 1. Characteristics of workshop participants.

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

n: number; IQR: Interquartile Range; HCT: hematopoietic stem cell transplantation. *Two nurses. **One nurse and one medical student. °Self-reported proficiency in English was reported using a Likert scale of 1 (not at all fluent) to 10 (extremely fluent). The number of $ symbols used indicates the number of missing participants.

ods 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 *Online Supplementary Methods* and *Online Supplementary Table S4* for details).

Statistical analysis

For details of the statistical analysis see the *Online Supplementary 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 (*Online Supplementary Table S1*).

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) (*Online Supplementary Table S1*). 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 six 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 score of 7 (IQR: 1; range: 2-10) on a Likert scale of 1 (not at all fluent) to 10 (extremely fluent).

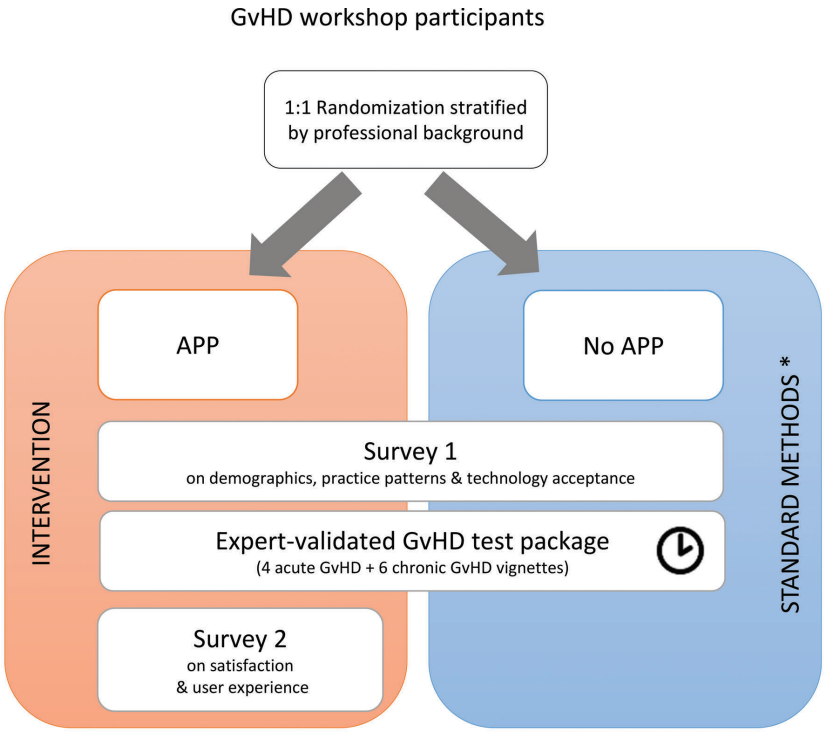

Figure 1. Study Design. APP: eGVHD App; GvHD: graft-*versus*-host disease. *E.g. own knowledge, 'fast facts' sheets, scoring sheets, standard operating procedures, copies of original guideline publications, or any other chosen resource.

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 1 to 10, IQR: 4, range 1-10), but performed with a relatively low level of confidence (median score of 5 on a Likert scale of 1 to 10, 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 *Online Supplementary Table S3*. 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 (*Online Supplementary Figure S1*) or professional background (*Online Supplementary Figure S2*). Similarly, neither the age of user (*Online Supplementary Figure S3*), the number of GvHD patients seen per week (*Online Supplementary Figure S4*), or self-reported comfort with using GvHD guidelines (*Online Supplementary Figure S5*) 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 Tables 4 and 5, 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 overes-

Table 2. Survey 1 results: pre-test practice patterns, technology access and technology acceptance data.

	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%)
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)$^{\$}$
Level of comfort ° - n (%)			
Low (≤ 4)	31 (42%)	17 (49%)	14 (35%)
Moderate (5-7)	35 (47%)	14 (40%)	21 (54%)
High (≥ 8)	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[11]	8 (10%)	3 (8%)	5 (13%)
The official NIH 2014 paper form[9]	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[11]	NA	NA	6 (15%)
The official NIH 2014 paper form[9]	NA	NA	9 (23%)
Other	NA	NA	7 (18%)
Not answered	NA	NA	1 (3%)
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 location 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 (0%)	1 (3%)
Not answered	2 (3%)	0 (0%)	2 (5%)
Predicted type of device* - n (%)			
Cellphone	43 (56%)	25 (68%)	18 (45%)
Tablet	5 (7%)	2 (5%)	3 (8%)
Laptop	6 (8%)	3 (8%)	3 (8%)

continued on the next page

continued from the previous page

Desktop	32 (42%)	10 (27%)	22 (55%)
Other	0 (0%)	1 (0%)	0 (0%)
Not answered	2 (3%)	0 (0%)	2 (5%)
Median importance of the availability of the app in my native language°	4 (IQR 5; range: 1-10)$$	4 (IQR 6; range: 1-10)	4 (IQR 5; range: 1-10)$$
Median reported level of likelihood of using the app°	8 (IQR 3; range: 1-10)$$$$	7.5 (IQR 3; range: 1-10)$	8 (IQR 4; range: 1-10)$$$$

n: number; IQR: Interquartile Range; NA: Not applicable. *Several answers were possible. °Reported on a Likert scale of 1 (lowest) to 10 (highest). The number of $ symbols used indicates the number of missing participants.

timation 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 2 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 (κ_{BP}= 0.46, 95%CI: 0.23-0.68) compared to the "No APP" group (κ_{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 *Online Supplementary Table S4*. 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) and 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 of '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 the time needed to score clinical vignettes, with an excess of approximately 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 health 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 presupposes 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 crosstalk 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]

Table 3. Graft-*versus*-host disease (GvHD) assessment accuracy and timing results.

Results for the complete GvHD test package (median)	APP (n=37)	No APP (n=40)
Correctly diagnosed vignettes	10 (IQR 1; range 5-10)	6.5 (IQR 3; range 2-9)
Correctly scored vignettes	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	4 (IQR 0; range 2-4)	2 (IQR 2; range 0-4)
Correctly scored chronic GvHD vignettes	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)

n: number; IQR: Interquartile Range; Std dev: standard deviation. The maximum number of correct answers for the whole package was 10 (4 for acute GvHD and 6 for chronic GvHD).

The lack of consensus and knowledge of the most recent guidelines was perhaps 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 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 the invitation to participate in 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 draw 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 specific response criteria,[18,19] which were not investigated in this project. Future studies, therefore, need to evaluate the use and impact of the eGVHD App in the clinic. This will also allow the evaluation of the App in situations where the patient does not present with GvHD, considering that the test package studied 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 stress, 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 10) is in line with the results of a previous validation study carried out in a more real-world setting. This study included actual patient examinations and showed that only 50-75% of freshly trained clinicians actually agreed with experts on the overall severity score of the evaluated chronic GvHD patients.[6] Mitchell *et al.* concluded that a single training session was not sufficient to achieve consistently acceptable inter-rater agreement between novice healthcare practitioners and GvHD experts. Clinical training in GvHD physical examinations may thus be necessary 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, aimed at making healthcare professionals ultimately independent of technological assistance.

The eGVHD App is currently limited to a calculator func-

Table 4. Detailed results of participants for graft-versus-*host* disease (GvHD) vignettes compared to the Expert Gold Standard - GvHD diagnosis.

Expert Gold Standard Diagnosis	Results from the "App" group given by 37 participants - n (%)						
	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%)

Expert Gold Standard Diagnosis	Results from the "No App" group given by 40 participants - n (%)						
	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%)

n: number; "Missing" corresponds to a lack of answer; "Other" corresponds to any answer not matching the proposed choices.The number of ° symbols used indicates the number of clinical vignettes involved.The highlighted diagonal corresponds to a perfect agreement between participants and expert results.

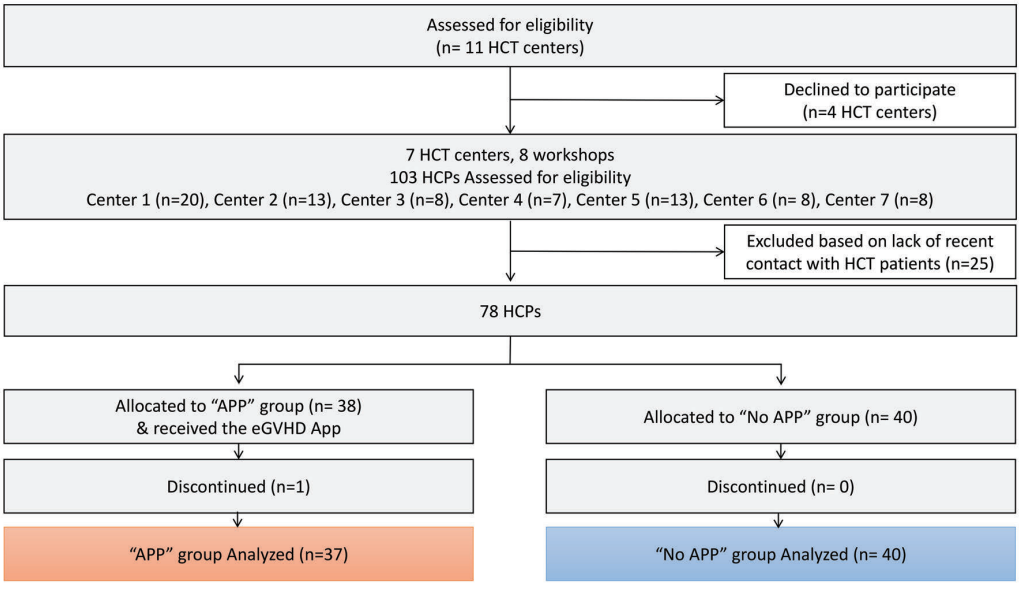

Figure 2. CONSORT flow diagram. APP: eGVHD App; HCPs: healthcare professionals; HCT: hematopoietic stem cell transplantation. n: number.

Table 5. Detailed results of participants for graft-*versus*-host disease (GvHD) vignettes compared to the Expert Gold Standard – GvHD Severity Scoring.

Expert Gold Standard Severity Scoring	Results from the "App" group given by 37 participants - n (%)									
	Grade II	Grade III	Grade IV	Grade	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%)

Expert Gold Standard Severity Scoring	Results from the "No App" group given by 40 participants - n (%)									
	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%)

n: number; "Missing" corresponds to a lack of answer; "Other" corresponds to any answer not matching the proposed choices. The number of ° symbols used indicates the number of clinical vignettes involved. The highlighted diagonal corresponds to a perfect agreement between participants and expert results.

tion 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 valid criticism expressed about the time-consuming aspect of its 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 assurance process at every new adaptation of the tool. This is not practically feasible in real life, and is probably, more often than not, unnecessary. Health regulation 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 widely available (*www.uzleuven.be/egvhd*) for all healthcare professionals who wish to obtain bedside user-friendly assistance in their GvHD assessment, and 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, the 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 individual patient into his or her 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 the practical issues.

Acknowledgments

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 PMID: 29872128].
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.