Title: Early diagnostic suggestions improve accuracy of family physicians: a randomized controlled trial in Greece.

Running head: Early diagnostic suggestions improve accuracy of Greek family physicians

Olga Kostopoulou, 1 Christos Lionis, 2 Agapi Angelaki, 2 Salma Ayis, 1 Stevo Durbaba, 1 Brendan C. Delaney 1

1 King’s College London, Department of Primary Care & Public Health Sciences, London, UK
2 University of Crete, Clinic of Social and Family Medicine, Heraklion, Greece

Correspondence: Dr Olga Kostopoulou, King’s College London, Faculty of Life Sciences & Medicine, Department of Primary Care & Public Health Sciences, Capital House, 42 Weston Street, London SE1 3QD, UK
E-mail: olga.kostopoulou@kcl.ac.uk
Tel: +44 (0)20 7848 6757
Fax: +44 (0)20 7848 6620

Word count: 2999
ABSTRACT

Background: In a recent RCT, providing UK family physicians with "early support" (possible diagnoses to consider before any information gathering) was associated with diagnosing hypothetical patients on computer more accurately than control. Another group of physicians, who gathered information, gave a diagnosis, and subsequently received a list of possible diagnoses to consider ("late support"), were no more accurate than control, despite being able to change their initial diagnoses.

Objective: To replicate the UK study findings in another country with a different primary healthcare system.

Methods: All study materials were translated into Greek. Greek family physicians were randomly allocated to one of three groups: control, early support and late support. Participants saw nine scenarios in random order. After reading some information about the patient and the reason for encounter, they requested more information to diagnose. The main outcome measure was diagnostic accuracy.

Results: 150 Greek family physicians participated. The early support group was more accurate than control (OR 1.67 [1.21 to 2.31]). Like their UK counterparts, physicians in the late support group rarely changed their initial diagnoses after receiving support. The pooled odds ratio for the early support vs. control comparison from the meta-analysis of the UK and Greek data was 1.40 [1.13 to 1.67].

Conclusion: Using the same methodology with a different sample of family physicians in a different country, we found that suggesting diagnoses to consider before physicians start gathering information was associated with more accurate diagnoses. This constitutes further supportive evidence of a generalizable effect of early support.

Keywords: diagnosis, diagnostic error, decision support systems, decision making
INTRODUCTION

Diagnostic errors are common.\textsuperscript{1-4} They can lead to considerable morbidity and mortality, and are the commonest reason for litigation against family physicians.\textsuperscript{5-7} Computerized diagnostic support systems (CDSS) have been proposed as one way of reducing the occurrence of diagnostic errors;\textsuperscript{8} nevertheless, they are not used in routine clinical practice, primarily because they do not integrate with electronic health record (EHR) systems, and also because the effect of providing support at different points in the diagnostic process is not well understood.

We hypothesized that providing diagnostic support early in the clinical encounter would improve diagnostic accuracy. It is well known that physicians generate diagnostic hypotheses early on, within seconds, and with minimal information.\textsuperscript{9,10} These hypotheses guide subsequent information search and interpretation.\textsuperscript{11,12} Influencing those initial hypotheses may therefore be instrumental in improving final accuracy. To test this hypothesis, we recently conducted a randomized controlled trial with UK family physicians diagnosing detailed patient cases via a web-link and concurrent telephone contact with a researcher.\textsuperscript{13} The trial tested a principle of “early” diagnostic support against control. Specifically, we provided one group of physicians with diagnoses to consider before they started gathering any information, based only on initial patient information (age, sex, risk factors, past medical history) and the reason for encounter. We found that this minimal support improved diagnostic accuracy, across a range of difficulty and irrespective of physician experience (years in family medicine). The 6% absolute difference in diagnostic accuracy obtained between the “early support” and control groups compares favorably with studies that tested fully developed CDSSs.\textsuperscript{14,15} Physicians in a “late support” group, who received diagnostic suggestions after completing their information gathering and giving their own diagnoses, were no better than control.
As part of the TRANSFoRm project (http://www.transformproject.eu), we aim to
design a computerized diagnostic support tool for family physicians across Europe.
We therefore needed to explore whether basic principles of support developed and
tested in one country could apply to other countries with different healthcare systems.
This is important, as widespread adoption of next generation CDSSs will rely on the
ability to integrate with a wide range of EHR systems and to demonstrate benefit in
Primary Care in different countries. It has already been shown that GPs in Malta, the
Netherlands, and Serbia see patients with similar reasons for encounter and make
similar diagnoses, so a prima facie case exists that a CDSS should be similarly
effective.\textsuperscript{16}

We therefore chose to test the principle of early diagnostic support in Greece, a
country with a primary care system very different from that of the UK. Primary care is
well established in the UK, as is computerization of all family practice clinics. All
patients have an electronic health record. Primary care in Greece is relatively new
and not well integrated in the healthcare system.\textsuperscript{17} Not all patients have an electronic
health record and not all family physicians use a computer for their consultations.\textsuperscript{18}
Diagnostic error has not received any attention in the country and family physicians
are not used to taking part in research. Clearly, replicating the UK study results in
such a different environment would increase our confidence in the potential of early
support to improve diagnostic accuracy across Europe.

There were 35,400 certified family physicians in England in 2011,\textsuperscript{19} while Greece
numbered only 1450 certified family physicians in 2007 (there are no numbers for
2011).\textsuperscript{20} Given the small number of family physicians in Greece and the lack of
culture of research participation, we did not expect to achieve an equivalent sample
size with the UK study (N=297). The university of Crete has a database of 383 Greek
family physicians, who have taken part in studies and educational events organized by the university. We decided to approach all physicians in the database, and explore whether the same direction of effect could be obtained in Greece as in the UK.

**METHODS**

**Materials and procedure**

To ensure an exact replication, we followed the same methods as in the UK study. All materials, including instructions, the patient scenarios and the diagnostic suggestions were translated to Greek and back-translated to English, checked by the UK team and any discrepancies were fed back to the Greek team who modified the materials accordingly. The researcher who collected the data in Greece (AA) underwent a two-day training session by the researcher who performed the data collection in the UK.

Data collection took place remotely over the Internet. The same web tool developed for the UK study was used. It randomized participants to groups according to a pre-determined blocked randomization sequence, randomized the scenario presentation order, presented the scenarios to participants, and collected the data. Participants were in simultaneous phone communication with the researcher who operated the site and guided them through the task during a single session.

Upon logging in on the study site, participants were automatically allocated to one of the three study groups, according to the blocked randomization sequence. Participants logged in at a pre-arranged date and time, so that the researcher could phone them at the start of the session. Once logged in, they filled in their demographic details: gender, whether they were family physicians or trainees in family medicine, year of qualification in family medicine, and whether they were or had been trainers of other physicians. They were then presented with a training
scenario and finally, with the 9 scenarios in random order. Each scenario was
designed to depict one correct diagnosis, rather than several possibilities. For more
information on the scenarios, see Kostopoulou et al.\textsuperscript{13} All patients had names, which
indicated their gender. At the start of each scenario, participants read some
introductory information about the patient: age, ethnicity, height, weight, BMI, alcohol
intake, smoking status, last blood pressure measurement, past medical history,
medications, last consultation, and a description of their appearance upon entering
the room. They also read the patient’s reason for encounter (RfE). For example:
“Good morning doctor. How are you today? I haven’t been well for about a week. I’ve
been off work with a really bad stomach pain and diarrhea. I thought that I had picked
up a bug and that it would get better by itself but it hasn’t.”

After confirming that they had read this initial information by clicking a button on the
screen, participants could start asking questions (history, examinations,
investigations) in order to find out what was wrong with the patient. The researcher
chose the appropriate answer to each question from a pool of answers (each
scenario had been extensively pre-tested and piloted to ensure that all expected
questions had an answer). If physicians asked a question for which an answer was
not predetermined, the researcher chose appropriately from a pool of generic
negative answers (“no”, “this is not a problem”, “I have not had this before”, “the
result is negative”, etc.). The participants read the answer, which was displayed on
their screen. Whenever they wished, they could finish the “consultation”, record the
diagnosis that they considered most likely and select one or more management
options from a list: refer, prescribe, arrange follow-up, give advice, and wait and see.
The control group could then proceed to the next scenario.

The “late support” group followed the same procedure as the control group, until they
submitted their diagnosis and management. They were then presented with a list of
diagnoses to consider. The list contained diagnoses that the participant had not yet discounted, given the information that he/she had gathered up to that point. The list thus differed for each participant, depending on his or her preceding information acquisition. For a detailed description of how this list was derived and updated according to each physician’s information search, see Kostopoulou et al. The participant could then choose to request more information, change diagnosis or keep the diagnosis already made. We recorded how often participants changed diagnoses following late support and if this impacted on accuracy.

The procedure for the “early support” group differed from that of the other two groups from the beginning of the task. Specifically, after participants confirmed that they had read the initial information (patient details and RfE), they were presented with a list of diagnoses to consider. This list was the same for every physician in the group, depending on the scenario, and was based on the specific patient details (age, gender, risk factors) and RfE. For a detailed description of how the early support lists were derived, see Kostopoulou et al. The list remained on the screen for at least 20 seconds and, after confirming that they had read it, participants could start asking questions. The list disappeared from the screen but participants could recall it at will by pressing a button. The list never changed during a scenario. In every other respect, participants proceeded through the task in the same way as the control group: they asked questions, received answers on their screen and, when ready, made a diagnosis, chose a management action and proceeded to the next patient. The ordering of the suggested diagnoses in both the early and late support lists was randomized.

**Recruitment**

We invited all family physicians and residents in family medicine for whom contact details were available in the database of the University of Crete. Specifically, we
invited 300 potential participants by e-mail, offering €100 recompense for participation. Data collection took place from February to December 2012.

Analyses
As in the UK study, we scored diagnosis as either correct or incorrect and management as either appropriate or inappropriate. Inappropriate management meant failing or delaying to deal with the condition, resulting in possible patient harm. All diagnoses and management decisions obtained in the study were scored by two family physicians according to pre-determined templates. We measured the effect of support (no support, early support, late support) on diagnostic accuracy using mixed effects logistic regression. We also explored the relationship between physician demographics (experience, gender, trainer or not) and diagnostic accuracy using mixed effects logistic regression.

We investigated the influence of support on information search (number of information requests and time taken) using mixed effects linear regression, and the influence of diagnostic accuracy on management, using mixed effects logistic regression. All regression models used random intercept to account for clustered data within participants, and scenario as a repeated measure.21

Finally we performed a meta-analysis using the odds ratios and CIs for the comparison between early support and control groups, pooling data from the UK and Greece.22 We used STATA 13.1 for all the analyses.

RESULTS
Of the 300 family physicians invited, 257 responded (86%) and 150 eventually participated in the study (50%). Participants had a median of 4 years in family
practice (mean 5.34, SD 5.35, range 0-23). Approximately half of the participants came from the Health Region of Crete; the rest came from the Health Regions of Central Macedonia, Peloponnese, Epirus, the Ionian Islands, and Western Greece. The sample consisted of 43.3% females (40.6% females in the 150 non-participants) and 20% residents in family medicine, reflecting the proportion of females and residents in the population of Greek family physicians (44% females, 20% residents). Table 1 presents demographics for both the Greek and UK samples of participating physicians per study group.

Mean diagnostic accuracy (proportion of correct diagnoses over all diagnoses) was 0.60 for control [95% CI 0.55 to 0.65], 0.71 for early support [0.67 to 0.76] and 0.69 for late support [0.65 to 0.73] (see Table 2 for a comparison with the UK results). Early support was associated with increased diagnostic accuracy: the odds of diagnosing correctly were 1.67 times higher than control (OR 1.67 [1.21 to 2.31], \( P = 0.002 \)).

The late support group was also more accurate than control (OR 1.50 [1.11 to 2.01], \( P = 0.008 \)). We explored this further by counting the number of times that an initial diagnosis was changed as a result of participants receiving late support (50 GPs diagnosing 9 scenarios = 450 occasions). We found that changes occurred only 4% of the time (for a comparison with changes in the UK group, see Table 3). In 7 instances, an initial diagnosis was changed from incorrect to correct. In all other instances, an initial incorrect diagnosis was changed to another incorrect diagnosis. There were no instances where an initial correct diagnosis was changed to an incorrect one. This indicated that the better performance of the late support group in comparison to control was not due to the support received. We ran a logistic regression on accuracy, this time using the initial diagnoses (i.e. before the support) rather than the final diagnoses of the late support group (we did not change the
scoring of accuracy for the other two groups, as their accuracy was only measured once, at the end of each scenario). We found that the late support group was significantly more accurate than control even before any support was provided (OR 1.41 [1.04 to 1.91], \( P = 0.027 \)).

To explain this, we explored potential differences in the demographics of the groups (Table 1). A Kruskall-Wallis test found no significant differences in experience (number of years in family practice) between groups (\( P = 0.27 \)). A logistic regression model with experience as the only predictor showed no association with diagnostic accuracy (\( P = 0.479 \)). Using experience as a covariate in the regression model that tested for the association between accuracy and support left the effect of late support almost unchanged (OR 1.44 [1.07 to 1.92], \( P = 0.015 \)). Chi-square tests found no significant differences between the study groups in the proportion of females (\( P = 0.30 \)) and trainers (\( P = 0.593 \)). Neither of these variables was associated with diagnostic accuracy. Adjustments for these in a multivariable model of accuracy on support did not change the results. We also explored the possibility that the late support group became more accurate from one scenario to the next simply by receiving the list of diagnostic suggestions in the first scenario. Including scenario sequence (i.e. first, second, etc.) as a predictor in the regression of diagnostic accuracy on support did not change the results of the model.

We did not detect any differences in terms of information search between groups, either in the number of requests for information (\( P = 0.58 \)) or in the time taken (\( P = 0.53 \)). Appropriateness of management was strongly associated with diagnostic accuracy (OR 11.93 [9.53 to 14.94], \( P < 0.001 \)).
The pooled odds ratio for the early support vs. control comparison from the meta-analysis of the UK and Greek data was 1.40 [1.13 to 1.67] (Figure 1). The test for heterogeneity was not significant ($P = 0.27$).

**DISCUSSION**

Summary of findings
This randomized controlled trial of a principle of diagnostic support that took place in Greece replicates the results from an earlier RCT conducted in the UK and thus, provides further supportive evidence for a generalizable effect of “early support”: in both countries, family physicians, who were provided with a list of diagnoses to consider before engaging in any information gathering, were significantly more accurate than an unaided control group. Both studies used the same, carefully constructed, patient scenarios presented to physicians on computer, and exactly the same methodology. In both countries, physicians who gave their own diagnosis first and then saw a list of diagnoses that could still not be discounted (the “late support” group), rarely changed their own diagnosis, even though they had the opportunity.

Strengths and limitations
The design of the study, which required physicians to record a diagnosis both before and after receiving late support, allowed us to measure baseline accuracy and thus avoid the erroneous conclusion that “late support improved accuracy”. The superior accuracy of the late support group over control was mainly due to their superior accuracy at baseline, i.e. before they received the support. We examined several potential confounders of the association between accuracy and support that could explain this, such as physician experience and gender. None were found to be associated with accuracy. Moreover, these variables did not change the association between accuracy and support in the sensitivity analyses. Therefore, the difference
at baseline between the two groups could be due either to chance, which can never be eliminated entirely with randomization. Indeed, a limitation of the Greek study is its small sample size, compared to the UK study. This could be responsible to some extent for some of the between-groups differences, and limits our ability directly to compare the effect sizes between the two countries. Nevertheless, a difference in the same direction between “early support” and control was found in both studies, increasing certainty that such a difference exists and is not a false positive.

Conclusion
Despite important differences between the healthcare systems of Greece and the UK, mean diagnostic accuracy in the two samples was very similar (Table 2). This provides a reassurance that our generic intervention to improve accuracy, designed with family physicians of one country in mind, will most likely benefit family physicians in other countries. This is extremely important when designing diagnostic support for primary care across Europe. It also constitutes further evidence for the similarities of the diagnostic task in family medicine between European countries.24

Family physician educators have suggested systematic ways of generating differential diagnoses as an aid to safe decision making.25 These strategies can be incorporated in a CDSS that presents physicians with diagnostic lists to consider early on in the encounter, ensuring, for example, that the most likely and the most serious diagnoses are shown first. In our study, the ordering of the suggested diagnoses was randomized; we aimed to test the effect of suggesting diagnostic alternatives either early or late in the diagnostic process versus control, and did not wish our findings to be limited to a specific presentation format. A CDSS employing the principle of early support would not only automate the generation of differential diagnoses, but also ensure their direct links to the best available evidence.26 We have now developed a prototype of such a CDSS and have integrated it with an
existing EHR system. Evaluation with family physicians is ongoing and will be completed during 2015.

Acknowledgements:
Drs Fotini Anastasiou, Elena Petelos and Ioanna Tsiligianni contributed to the translation of the study materials and the scoring of the management decisions. Andrea Rosen, research assistant for the UK study, prepared some of the data for analyses and liaised with the Greek team during the study. Drs Tom Round and Ellen Wright scored the accuracy of the participants’ diagnoses.

DECLARATION
Funding: This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 247787 [TRANSFoRm], and from the NIHR Biomedical Research Centre at Guy’s and St Thomas’ NHS Foundation Trust and King’s College London.

Ethical approvals were obtained from the following hospitals:
The University General Hospital of Heraklion (protocol number 12845/22/11/2010)
The General Hospital ‘Venizeleio – Pananeio’ of Heraklion (protocol number 32-12/5/10-5-2011)
The General Hospital ‘Agios George’ of Chania (protocol number 10999/24/06/2011)
The General Hospital – Health Care Center of Siteia (protocol number 1512/184-2011)
The General Hospital ‘Agios Paulos’ of Thessaloniki (protocol number 102/13-4-2011)
The General Hospital ‘Papageorgiou’ of Thessaloniki (protocol number 181/8-4-2011)
The General Hospital ‘G. Chatzicosta’ of Ioannina (protocol number 7/27-4-2011)

**Disclosures:** Professor Lionis is Associate Editor of the journal. The authors have no conflict of interests.
References

Figure 1. Forest plot of the odds ratio meta-analysis of the UK and Greek data. Odds ratios refer to the comparison between control and early support groups in terms of diagnostic accuracy.
Table 1. Physician demographics by study group in Greece and the UK.

<table>
<thead>
<tr>
<th>GREECE</th>
<th>Control (n=51)</th>
<th>Early (n=49)</th>
<th>Late (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(median &amp; range)</td>
<td>3 years (0-12)</td>
<td>4 years (0-22)</td>
<td>5.5 years (0-23)</td>
</tr>
<tr>
<td>% females</td>
<td>37</td>
<td>41</td>
<td>52</td>
</tr>
<tr>
<td>% trainers</td>
<td>18</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>UK</td>
<td>Control (n=99)</td>
<td>Early (n=99)</td>
<td>Late (n=99)</td>
</tr>
<tr>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(median &amp; range)</td>
<td>3 years (0-33)</td>
<td>5 years (0-34)</td>
<td>7 years (0-34)</td>
</tr>
<tr>
<td>% females</td>
<td>60</td>
<td>50</td>
<td>52</td>
</tr>
<tr>
<td>% trainers</td>
<td>14</td>
<td>17</td>
<td>13</td>
</tr>
</tbody>
</table>
Table 2: Mean diagnostic accuracy (mean proportion of correct diagnoses over all diagnoses across the 9 scenarios) and 95% CIs for the Greek and UK samples of family physicians. Kostopoulou et al. report on the UK sample and findings.\textsuperscript{13}

<table>
<thead>
<tr>
<th></th>
<th>Greece (n=150 family physicians)</th>
<th>UK (n=297 family physicians)\textsuperscript{13}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control group (n=51)</td>
<td>Control group (n=99)</td>
</tr>
<tr>
<td></td>
<td>0.60 [0.55 to 0.65]</td>
<td>0.63 [0.60-0.67]</td>
</tr>
<tr>
<td>Early support group</td>
<td>0.71 [0.67 to 0.76]</td>
<td>Early support group (n=99)</td>
</tr>
<tr>
<td>(n=49)</td>
<td></td>
<td>0.69 [0.66-0.73]</td>
</tr>
<tr>
<td>Late support group</td>
<td>0.69 [0.65 to 0.73]</td>
<td>Late support group (n=99)</td>
</tr>
<tr>
<td>(n=50)</td>
<td></td>
<td>0.65 [0.62-0.70]</td>
</tr>
<tr>
<td>Total</td>
<td>0.67 [0.64-0.70]</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.66 [0.64-0.68]</td>
</tr>
</tbody>
</table>
Table 3. Number of times (and %) that the late support group changed their initial diagnosis, after receiving “late support” (i.e., an individualized list of diagnostic suggestions after a physician gave a preliminary diagnosis and management). The table presents data from the Greek (n=150) and UK (n=297) studies. In the Greek study, % changes are based on 50 physicians in the late support group diagnosing 9 scenarios (450 diagnostic instances). In the UK study, % changes are based on 99 physicians in the late support group diagnosing the same 9 scenarios (891 diagnostic instances).

<table>
<thead>
<tr>
<th>Changes of the initial diagnosis</th>
<th>Late support group, Greek study</th>
<th>Late support group, UK study(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>From incorrect to correct</td>
<td>7 (1.6%)</td>
<td>33 (3.7%)</td>
</tr>
<tr>
<td>From incorrect to incorrect</td>
<td>10 (2.2%)</td>
<td>24 (2.7%)</td>
</tr>
<tr>
<td>From correct to incorrect</td>
<td>0</td>
<td>1 (0.1%)</td>
</tr>
<tr>
<td>Total changes</td>
<td>17 (3.8%)</td>
<td>58 (6.5%)</td>
</tr>
</tbody>
</table>