

# 一项“有效”的健康干预，为什么换个合理模型就不再有效？

## 印度饮水砷信息试验的独立再分析

Everion Cao

GHEAS，加拿大安大略省滑铁卢

ORCID：0009-0005-8238-8583；通信邮箱：everioncao@gmail.com

版本：讨论稿 1.1，2026 年 7 月 22 日

欢迎评论：本文尚未投稿，欢迎对数据处理、模型选择和结论提出批评与建议

### 摘要

#### 研究背景

印度比哈尔邦的一项随机试验曾报告：向居民提供饮水砷检测结果和相关知识，可以减少不安全饮水，并可能改善健康。我们重新分析这项研究，并不是为了判断原作者“对”还是“错”，而是想回答一个更普遍的问题：同一批数据，在采用不同但都说得通的分析方法时，结论会不会发生根本变化？

#### 研究方法

我们使用公开的 1.2 版数据和代码。数据审计发现，部分村庄编号发生了可以确定的错位，另有三个村庄各存在一户分组记录无法确认。主分析修复了可以确定的错位，并排除这三个村庄，共保留 2,289 户、153 个村庄。我们先完整复现原论文的模型，再使用更直接对应随机分组的模型重新估计干预效果。后者不控制“知识是否提高”等干预以后才发生的变量。我们还分析了基线水砷浓度与胃部问题、抑郁和皮肤问题之间的关系。

#### 主要结果

原论文的健康结果可以复现：小组干预与胃部问题减少 9.34 个百分点相关， $p=0.005$ 。但在不控制干预后知识变化的主模型中，这一估计缩小到 4.36 个百分点，95% 置信区间为 -11.00 至 2.27，无法排除没有效果的可能。个体干预的结果也从 -4.16 个百分点变为 +2.83 个百分点。模型分解表明，结论的转折主要发生在加入“知识变化”之后。

个体干预组的平均水砷从 27.4  $\mu\text{g/L}$  降到 24.2  $\mu\text{g/L}$ ，但相对于对照组的调整后差异为 -5.11  $\mu\text{g/L}$  (95% CI -11.99 至 1.76)，仍不够确定。基线水砷越高，胃部问题越多；但水砷与抑郁或皮肤问题都没有稳定关系。皮肤的复杂曲线模型一度给出极小的  $p$  值，但进一步检查发现，这个结果完全由一条 1.0  $\text{mg/L}$  的极端记录推动。去掉这一条记录后， $p$  值从约  $2.0 \times 10^{-7}$  变为 0.14。

#### 结论

这项干预可能提高了居民对砷的认识，也可能改变了部分家庭的饮水选择；但公开数据尚不足以证明水砷得到稳定下降，更不足以证明砷暴露降低带来了健康改善。这个案例提醒我们：代码能够复现，不等于结论已经可靠；一个统计上显著的结果，也可能高度依赖模型中放入了什么变量。

关键词：饮水砷；随机试验；干预后变量；剂量—反应；皮肤损害；可复现研究

# 1. 为什么重新分析这项研究？

砷是全球饮用水中最重要的化学污染物之一。世界卫生组织（WHO）将 10 μg/L 作为饮水指导值，并建议在现实条件允许时尽量降得更低[2]。印度的饮水标准也采用 10 μg/L[5]。不过，这条线是管理标准，不是“低于它就绝对安全、高于它就一定生病”的生物学界线。

长期摄入无机砷，最典型的表现是皮肤色素改变、掌跖角化和皮肤癌；此外还可能影响神经、心血管、代谢和发育，并增加膀胱癌和肺癌风险[2-4]。胃痛、呕吐和腹泻更常见于急性高剂量中毒。慢性暴露当然也可能伴随胃肠不适，但“胃炎、便秘或腹泻”非常常见，单凭这些症状很难判断是不是砷造成的。

Priyam、Salicath 和 Sutter 在 156 个村庄、2,334 户家庭中进行了三组随机试验。两个干预组都接受了砷知识介绍、家庭水样检测和结果告知；对照组观看无关视频，也不知道水样检测结果。论文认为，干预提高了饮用安全水的概率，并发现小组干预可能减少胃部问题[1]。

这项研究很适合作为方法学案例。它有公开数据、公开代码和随机试验设计，但也同时存在编码异常、干预后变量、非特异性健康结局以及多种合理模型。我们关心的不是挑出一个最漂亮的数字，而是看不同选择怎样一步步改变结论。

# 2. 我们怎样处理数据和模型？

## 2.1 公开资料与研究设计

基线调查在 2019 年 12 月至 2020 年 2 月进行，约一年后随访。公开资料来自 Edmond 数据集 1.2 版，包括面板数据、宽表、健康数据和 Stata 代码[1,6]。水砷由现场试剂盒检测，记录为 0、10、25、50、100、200、300、500 或 1,000 μg/L。健康状况由家庭自报。

## 2.2 数据中发现了什么问题？

在查看任何 GHEAS 效应结果之前，我们先核对不同文件之间的村庄、分组和基线信息。审计发现 22 条村庄编号错位，可以根据公开文件中的重复信息唯一修复。修复后，村庄 2016、2032 和 4022 仍各有一户的分组记录与同村其余 14 户不一致，公开资料不足以判断哪一个值才是真值。

因此，我们没有猜测，也没有请原作者替我们补齐答案。主分析整村排除这三个存在冲突的村庄，同时保留三种敏感性分析，检查不同处理方式是否会改变结论。

| 情景 | 怎样处理数据               | 用途    | 健康记录/村庄     |
|----|----------------------|-------|-------------|
| D0 | 修复 22 条错位；排除 3 个冲突村庄 | 主分析   | 2,289 / 153 |
| D1 | 修复错位；只排除 3 条冲突记录     | 敏感性分析 | 2,331 / 156 |
| D2 | 修复错位；按同村多数分组填入 3 条记录 | 敏感性分析 | 2,334 / 156 |
| D3 | 完全保留公开编码             | 复现原论文 | 2,334 / 156 |

主健康模型最终使用 2,232 条信息完整的 D0 记录。所有修改都有单独账本，原始压缩包没有被改动。

## 2.3 三个健康指标分别代表什么？

公开文件中可以直接分析的健康指标只有三个：家庭中是否有人出现胃部问题，家庭抑郁指标，以及家庭中是否有人出现皮肤问题。胃部问题包括胃炎、便秘或腹泻；抑郁指标来自 PHQ-9；皮肤问题包括皮肤刺激、黑变病或角化过度[1]。

这三个指标不能等量齐观。抑郁不是周围神经病变、麻木、认知或儿童神经发育的替代指标。皮肤问题与慢性砷中毒更接近，但它把非特异性的皮肤刺激与较典型的黑变病、角化过度合在一起，而且没有医生检查。胃部问题最常见，却也最不特异。

## 2.4 为什么要比较两类模型？

随机试验首先回答的是：被分到干预组，整体上带来了什么变化？因此，主模型只使用随机分组和随机化区组，不控制干预以后才出现的知识变化、交谈或水源选择。

原论文的健康模型加入了“知识变化”。但知识提高正是干预希望产生的结果之一。如果在回归中把它固定住，问题就从“干预总体有没有用”变成了“在知识变化相同的人之间，干预还有没有额外作用”。这不是同一个问题，而且可能引入新的偏差[8]。

所有二元结局都用线性概率模型表示，因此系数可以直接读作百分点差异。标准误按村庄聚类。我们还使用随机化推断、多重比较校正、逐村删除和四种数据情景来检验结果是否稳定。

水砷与健康的分析属于观察性关联，不能因为它出现在随机试验数据里就自动变成因果关系。我们比较了连续浓度、10  $\mu\text{g/L}$  阈值、三个浓度区间和自然样条曲线。

## 3. 我们发现了什么？

### 3.1 干预前后，水砷到底降了多少？

三个组在基线时的平均水砷都约为 27  $\mu\text{g/L}$ ，中位数都是 10  $\mu\text{g/L}$ 。主分析样本中，34.3% 的家庭超过 10  $\mu\text{g/L}$ ，18.8% 达到或超过 50  $\mu\text{g/L}$ 。

| 组别   | 基线均值 $\mu\text{g/L}$ | 期末均值 $\mu\text{g/L}$ | 同一家庭平均变化 $\mu\text{g/L}$ | >10 $\mu\text{g/L}$ 基线→期末 | ≥50 $\mu\text{g/L}$ 基线→期末 |
|------|----------------------|----------------------|--------------------------|---------------------------|---------------------------|
| 对照   | 26.6                 | 27.1                 | +0.08                    | 27.3%→30.1%               | 15.3%→15.0%               |
| 个体干预 | 27.4                 | 24.2                 | -3.47                    | 39.5%→36.1%               | 19.6%→15.8%               |
| 小组干预 | 26.8                 | 26.7                 | +0.02                    | 35.7%→36.4%               | 21.4%→16.8%               |
| 全体   | 26.9                 | 26.0                 | —                        | 34.3%→34.3%               | 18.8%→15.9%               |

个体干预组看起来下降了一些，但基线调整后，相对于对照组的差异为 -5.11  $\mu\text{g/L}$  ( 95% CI -11.99 至 1.76 ;  $p=0.144$  )。这个区间既容许有一定改善，也容许几乎没有改善。小组干预的差异为 -0.37  $\mu\text{g/L}$  ( $p=0.909$ )，没有看到下降信号。个体干预组期末平均浓度仍是 10  $\mu\text{g/L}$  指导值的约 2.4 倍，而且 36.1% 的家庭仍超过指导值。

### 3.2 原论文关于“不安全饮水”的结果能否复现？

能。完全按照公开文件和原模型分析，个体干预的“分组×期末”估计为 -7.28 个百分点 ( $p=0.044$ )，小组干预为 -2.36 个百分点 ( $p=0.445$ )，几乎等于论文表格中的 -7.3 和 -2.4[1]。

但这个结果处在边缘位置。不加入干预后的知识和交谈变量时，个体干预估计为 -6.00 个百分点， $p=0.060$ ；加入知识后才变为 -7.33， $p=0.043$ 。在结构修正后的 D0 样本中，完整原模型得到 -7.11， $p=0.055$ 。原论文也说明，经过 Holm-Bonferroni 多重比较校正后，个体干预结果不再显著。

因此，更贴近数据的说法是：个体干预可能减少了超过 10  $\mu\text{g/L}$  的家庭比例，但证据较弱，并且会随模型和数据处理方式变化；小组干预没有显示清楚的水砷改善。

### 3.3 为什么胃部健康结论发生了变化？

| 结局与模型           | 个体干预，百分点 ( 95% CI )    | 小组干预，百分点 ( 95% CI )     | 怎样理解          |
|-----------------|------------------------|-------------------------|---------------|
| 胃部，原论文模型 D3     | -4.16 ( -11.29, 2.97 ) | -9.34 ( -15.85, -2.84 ) | 可以复现；控制了干预后变量 |
| 胃部，GHEAS 主模型 D0 | +2.83 ( -3.65, 9.30 )  | -4.36 ( -11.00, 2.27 )  | 更接近随机分组的总体效果  |
| 抑郁，GHEAS 主模型 D0 | -2.89 ( -10.88, 5.10 ) | -0.08 ( -7.83, 7.68 )   | 没有明确效果        |
| 皮肤，GHEAS 主模型 D0 | +1.08 ( -4.07, 6.23 )  | -2.24 ( -6.84, 2.36 )   | 没有明确效果        |

六个健康比较经过 Holm 校正后都不能排除零效应。胃部主模型的随机化推断  $p$  值分别为 0.419 和 0.208。

# 模型调整改变表观健康效应

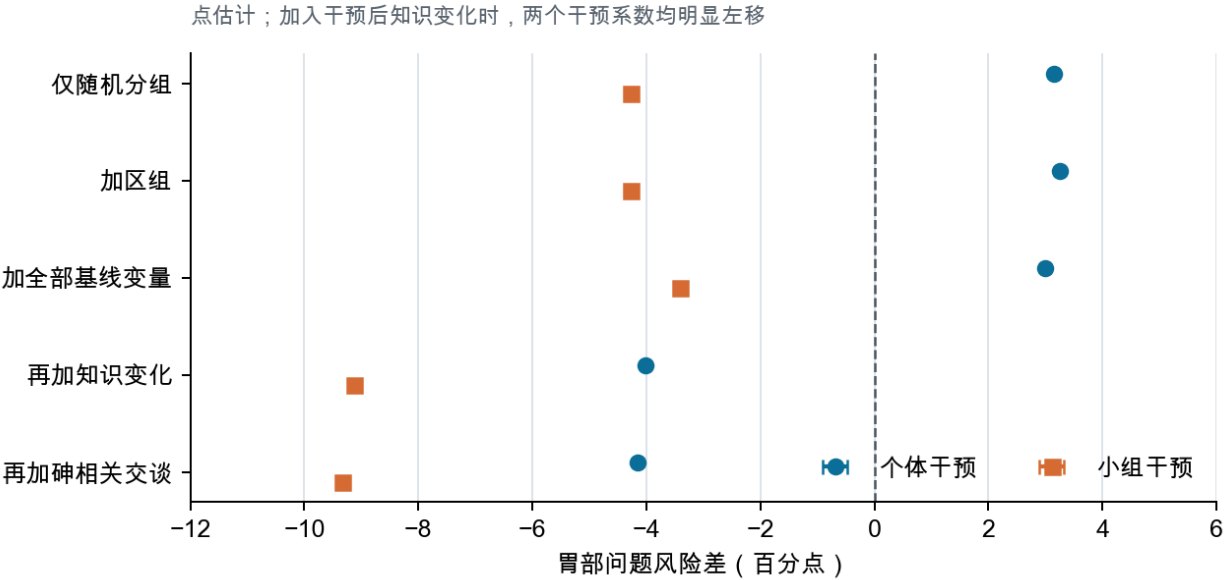


图 1. 干预后知识调整与胃部问题效应估计的变化。

图 1 展示了结论如何一步步改变。只加入基线变量时，个体和小组干预的胃部估计仍为 +2.99 和 -3.41 个百分点。加入干预后的知识变化后，它们立刻变为 -4.02 和 -9.13；再加入砷相关交谈，变化很小。也就是说，原论文中显著的小组健康结果，主要是在模型控制“知识变化”之后出现的，而不是因为控制了基线差异。

## 3.4 水砷与抑郁、皮肤问题有关系吗？

| 健康指标 | 每增加 10 $\mu\text{g/L}$ ，百分点 ( 95% CI ) | $>10 \mu\text{g/L}$ ，百分点 ( 95% CI ) | 10- $<50$ vs $\leq 10 \mu\text{g/L}$ | $\geq 50$ vs $\leq 10 \mu\text{g/L}$ |
|------|----------------------------------------|-------------------------------------|--------------------------------------|--------------------------------------|
| 胃部问题 | +0.43 ( 0.09, 0.78 )                   | +5.27 ( 0.86, 9.68 )                | —                                    | —                                    |
| 抑郁   | +0.03 ( -0.46, 0.52 )                  | +3.48 ( -2.04, 9.00 )               | +3.72 ( -2.60, 10.04 )               | +3.26 ( -3.68, 10.19 )               |
| 皮肤问题 | +0.07 ( -0.35, 0.49 )                  | -0.74 ( -5.11, 3.64 )               | -0.50 ( -6.13, 5.13 )                | -0.96 ( -6.00, 4.09 )                |

基线水砷每增加 10  $\mu\text{g/L}$ ，胃部问题增加约 0.43 个百分点， $p=0.015$ 。但抑郁和皮肤问题的结果都跨过零，而且无论把浓度看作连续数值、是否超过 10  $\mu\text{g/L}$ ，还是分成三个区间，结论都没有改变。逐条删除后，两个线性结果甚至都可以改变方向。

## 水砷与健康结局的关联

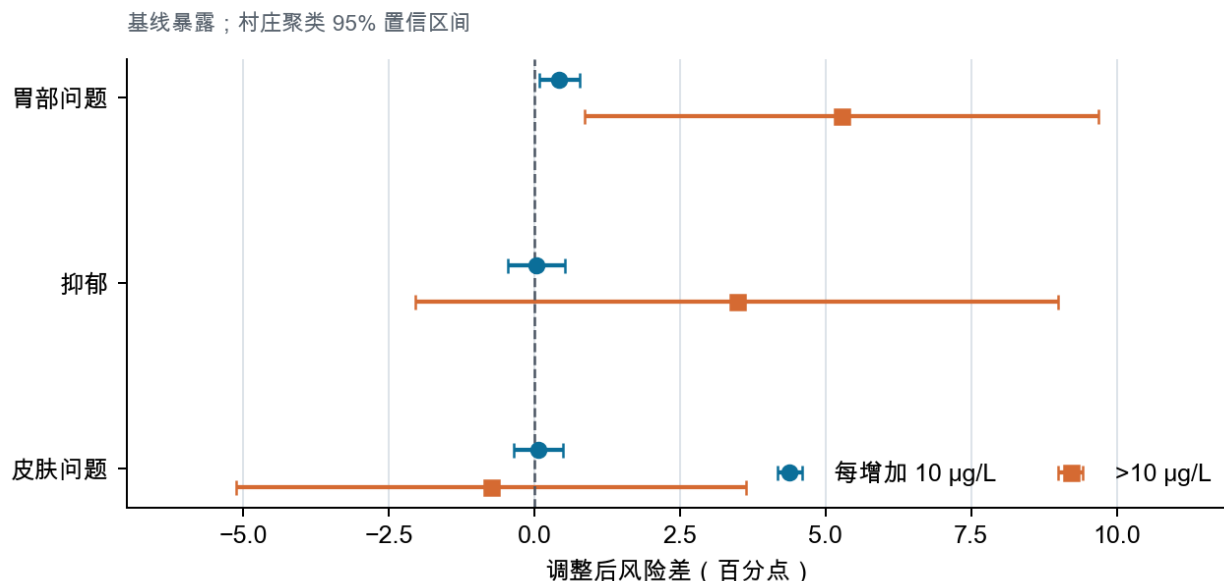


图 2. 基线水砷与三个家庭健康指标的调整关联。

皮肤的自然样条模型曾给出  $p \approx 2.0 \times 10^{-7}$ ，看上去非常“显著”。但  $1.0 \text{ mg/L}$  这一浓度只有一条记录，而且该户恰好报告了皮肤问题。只去掉这一条记录， $p$  值就变成 0.14；把最高浓度限制在  $0.3 \text{ mg/L}$  时， $p=0.746$ 。这个结果不是可信的生物学曲线，而是稀疏尾部与复杂模型共同造成的统计假象。

### 3.5 为什么三个健康指标给不出同一个故事？

胃部问题在完整样本中的报告率高达 72.0%，皮肤问题为 14.6%。胃部指标可能捕获了广泛的家庭症状，其中许多与砷无关。它确实与基线水砷呈正相关，但我们没有看到相应的皮肤模式，也没有神经系统指标来相互印证，更没有形成“干预降低水砷，水砷降低又改善健康”的完整证据链。

皮肤指标在病理上更贴近慢性砷危害，却有两个现实限制：没有临床确认，而且随访只有约一年。WHO指出，典型慢性砷性皮肤改变通常至少需要多年才出现[2]。因此，皮肤结果为阴性，既不能说明砷没有危害，也不能说明干预一定无效；更可能的解释是，这个指标和随访时间不适合观察慢性皮肤改善。

## 4. 这些结果意味着什么？

### 4.1 可以复现，不代表可以照原样解释

原论文的核心系数几乎都能复现，所以问题不是程序算错了。真正的分歧在于：模型回答了什么问题，控制了哪些变量，以及健康指标能否代表砷的真实危害。

如果目标是估计随机分组的总体效果，就不应把干预产生的知识变化当作普通背景变量控制。控制它以后得到的结果可以作为机制探索，但不能替代主要的意向治疗结果。在本案例中，正是这一步使健康系数翻转，并产生了显著结果。

### 4.2 胃部问题不能单独承担“砷改善健康”的证据

在以  $10\text{--}50 \text{ µg/L}$  为主的慢性暴露环境中，一般胃炎、便秘和腹泻并不足以识别砷病因。微生物污染、饮食、营养、卫生条件、感染和报告行为都可能改变这些症状。原论文自己也提出，胃部变化可能来自居民煮沸地表水后减少了细菌污染，而不一定来自水砷下降[1]。

要检验砷相关健康改善，今后的研究至少需要重复的实验室水样检测、尿砷等生物标志物、临床皮肤检查、周围神经检查、基线健康资料和更长时间的随访。

### 4.3 对数据分析方法有什么启示？

这个案例给出了几条很具体的经验：先核对数据单位和随机分组，再计算效果；把随机化总体效果与机制分析分开；如果要声称污染物改善了健康，必须先证明污染物暴露确实下降；选择健康指标时要考虑疾病特异性和潜伏期；复杂曲线出现惊人结果时，应先检查是不是少数极端记录造成；最后，不要让  $p < 0.05$  自动升级为强结论。

## 5. 研究的优势与限制

这项再分析保留了原始文件，记录了每一步修复，比较了四种数据处理情景，并且同时报告支持与不支持原结论的结果。我们还复现了原模型，使用随机化推断、多重比较校正和逐村影响分析，并明确标出哪些扩展分析是在看到前期结果后才提出的。

限制同样重要。公开健康文件没有家庭 ID，三个分组冲突无法知道真值；现场试剂盒只有离散读数，没有实验室复测、批次和操作人员信息；健康结局均为家庭自报，没有基线健康和个人对应关系；不同村庄之间是否共用水源、是否相互传播信息也无法确认。水砷与抑郁、皮肤的扩展分析属于结果可见后的探索，不是独立的验证性研究。

## 6. 结论

这项印度饮水水砷信息试验的公开结果具有很高的数值可复现性，但健康结论并不稳固。不控制干预后变量时，我们没有发现明确的胃部、抑郁或皮肤改善。个体干预可能减少了超过 10  $\mu\text{g/L}$  的家庭比例，但证据较弱；连续浓度的下降也不确定。

目前最合理的结论是：干预可能提高了知识，并改变了部分饮水行为；但公开证据还不能证明水砷得到稳定降低，也不能证明砷暴露降低改善了健康。这并不否定原研究的价值。相反，它说明一项研究最有价值的部分，有时不是一个显著的健康系数，而是它为我们提供了检验分析方法、发现证据边界的机会。

## 数据、代码与透明度声明

原作者通过 Edmond 以 CC BY 4.0 许可公开了 1.2 版数据，DOI 为 10.17617/3.LPX8BT[6]。GHEAS 保存了研究方案、异常处理规则、转换账本、数据情景、模型清单和分析结果。原始 ZIP 文件没有被修改，也没有联系原作者补填或解释公开记录。

Everion Cao 决定研究问题、分析原则和最终判断。OpenAI Codex ( 项目中称为 Jace ) 协助编程、执行和文字起草，但不得自行提高因果结论的强度，也没有用生成式 AI 填补缺失结局或决定冲突记录的真值。

## 利益冲突

作者声明不存在利益冲突。

## 参考文献

- [1] Priyam S, Salicath D, Sutter M. An information intervention in India reduces consumption of arsenic-contaminated drinking water and has small effects on health. PNAS Nexus. 2026;5(4):pgag093. <https://doi.org/10.1093/pnasnexus/pgag093>
- [2] World Health Organization. Arsenic: fact sheet. 7 December 2022. <https://www.who.int/news-room/fact-sheets/detail/arsenic>
- [3] World Health Organization. Exposure to arsenic: a major public health concern. 2019. WHO/CED/PHE/EPE/19.4.1. <https://www.who.int/publications/i/item/WHO-CED-PHE-EPE-19.4.1>
- [4] United States Environmental Protection Agency. IRIS assessment: inorganic arsenic, CASRN 7440-38-2. Updated 13 January 2025. [https://iris.epa.gov/ChemicalLanding/&substance\\_nmbr=278](https://iris.epa.gov/ChemicalLanding/&substance_nmbr=278)
- [5] Government of India, Ministry of Jal Shakti. Monitoring of contamination of ground water; Annex I, IS 10500:2012 drinking-water limits. 21 July 2022. <https://www.pib.gov.in/Pressreleaseshare.aspx?PRID=1843385>

- [6] Priyam S, Salicath D, Sutter M. Replication data for "An information-intervention in India reduces consumption of arsenic-contaminated drinking water and has small effects on health." Edmond, version 1.2. 2026. <https://doi.org/10.17617/3.LPX8BT>
- [7] George CM, Zheng Y, Graziano JH, et al. Evaluation of an arsenic test kit for rapid well screening in Bangladesh. *Environmental Science & Technology*. 2012;46:11213-11219.
- [8] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20(4):488-495. <https://doi.org/10.1097/EDE.0b013e3181a819a1>
- [9] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65-70.
- [10] Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief Patient Health Questionnaire mood scale (PHQ-9) in the general population. *General Hospital Psychiatry*. 2006;28:71-77.