Greedy and Exact Algorithms for Invitation Planning in Cancer Screening

Marco Gavanelli[1], Michela Milano[2], Sergio Storari[1], Luca Tagliavini[1], Paola Baldazzi[3], Marilena Manfredi[3], and Gianfranco Valastro[4]

[1] Department of Engineering - University of Ferrara
Via Saragat, 1 – 44100 – Ferrara, Italy
[marco.gavanelli,sergio.storari,luca.tagliavini]@unife.it
[2] DEIS - University of Bologna
Viale Risorgimento, 2 – 40136 – Bologna, Italy
mmilano@deis.unibo.it
[3] Department of Public Health - Sanitary Agency of Bologna
Via Montebello, 6 – 40136 – Bologna, Italy
[paola.baldazzi,marilena.manfredi]@ausl.bo.it
[4] NOEMALIFE SpA
Via Gobetti, 52 – 40129 – Bologna, Italy
gvalastro@noemalife.com

Abstract. Cancer screening is a method of preventing cancer by early detecting and treating abnormalities. One of the most critical screening phase is invitation planning since screening resources are limited and there are many people to invite. For this reason, smart resource allocation approaches are needed.
In the paper, we propose and compare two solutions for smart invitation plan definition, one based on greedy approaches and one based on Constraint Programming techniques that enable the definition of the optimal invitation plan.

1 Introduction

Cancer screening is a process finalized to the prevention of the illness from its starting phases. Early diagnosis of tumors is fundamental, because a timely intervention makes the healing easier and reduces the risk of death. In fact, uterus cancer, breast cancer and many other tumors are preventable and fully curable if they are early diagnosed.

We focused on cervical cancer screening, that enables the identification of tumors in the cervix. To reduce the mortality related to this kind of tumor it is necessary to ensure periodically pap-test screening [2] for the entire female population with age between 25 and 64 years.

The screening process is managed by the screening center manager and consists of several phases. First of all, the involved patients (composing the so-called *target population*) are identified, by excluding, e.g., people that have already a cancer and residents in other areas. Given the target population, the next step is to create an invitation plan for the screening examination. This plan should be coherent with the time availability of the centers in which the screening examinations are performed. Once this planning phase is finished, invitations are sent by mail to the target population, and usually patients are

visited at the screening center in the scheduled time slots. The screening process then proceeds in different ways, depending on the result of the screening examination.

Among all these phases, the most complex one, from an organizational point of view, is the invitation planning since the number of patients is typically high and the pap-test center resources (time and personnel) are limited. For this reason, we need smart resource allocation approaches, exploiting optimization technology.

Baker and Atherill [1] study, by means of simulations of queue theory, the order of patients to be invited. The order is then optimized by means of a sort of hill climbing algorithm; the objective is to minimize the dissatisfaction of the patients (modelled as a function of the waiting time), and the server idle time. Other authors [3, 5] use a weighted sum of the patients average waiting time and server idle time. In [1], authors analyzed datasets of pap-test invitations in order to identify probabilistic models of patient attendance and appointment rebooking.

In this paper, we describe the research activity carried on within an industrial project of the Emilia Romagna region of Italy for handling the invitations of cervix cancer screening in the Bologna district. In particular, we explain how invitation plans are currently generated and we propose two solutions to improve the efficiency of the process. The first is based on greedy algorithms: we show two algorithms and the corresponding results. The second is based on Constraint Programming techniques that provide the optimal invitation plan.

Performance evaluations have been conducted on exact and heuristic solutions by means of simulations on different scenarios involving different groups of women and different pap-test center resources.

2 Pap-test invitation management

The definition of the pap-test invitation plan is a very complex task since it involves many women and consequently requires a lot of resources. Involved information include: pap-test center resources, last pap-test examination date, screening history of each woman in the target population, women addresses.

Pap-test center resources are represented by time periods offered each day for the execution of the examinations. These time periods can change each month so each center regularly communicates its monthly agenda to the screening center manager. Usually the time assigned for each pap-test execution is 10 minutes. For this reason, given a pap-test invitation at time T, the next one can be scheduled at time T plus 10 minutes. Moreover, given the pap-test duration, the number of patients that can be invited a day D in a pap-test center is the number of 10 minute slots contained in its available time period.

The last pap-test examination date is important because the next expected pap-test should be performed three years after the last one.

The screening history of a woman is the collection of all the events happened during her screening process (e.g., received/refused invitations, pap-test results). Depending on these events, the woman is classified into three main priority levels: High Priority (HP), Normal Priority (NP) and Low Priority (LP). A woman is classified as High Priority when during her screening history a high risk event has occurred (e.g., if a tumor

was found and treated in the last two years). Normal Priority is associated to women that have accepted the last pap-test invitation and results were normal. Low priority is assigned to women who have not accepted the last pap-test invitation. The screening protocol prescribes to track such women and retry the pap-test invitation several times. Statistics show that LP women have very low probability of accepting an invitation: typically less than 30% of the invited women show up. For this reason, overbooking is a common practice: in our instance, the examination duration is reduced down to 3 minutes for LP patients. The assigned priority is one of the most important parameters for the definition of the pap-test invitation plan because usually a fixed percentage of the time-periods available in the pap-test centers is allocated for each priority level.

During the round, centers might be early or late on calls. In the first case, the center can be excluded from the invitation plan. In the second case, overbooking is performed.

The address of a woman is important because she should be invited in the nearest pap-test center in order to increase the probability of showing up.

Given the information described above, the definition of the monthly invitation plan is made in several steps. The screening center manager receives from the pap-test center the availability agenda for the next month expressed in minutes. A list of women to invite is identified by filtering the target population by choosing among the target population only those women whose invitation expires before a certain deadline. The overall time availability is subdivided in slots of 10 minutes each. A percentage of slots is then assigned to each priority level (default percentages are: 50% for High priority, 30 for Normal priority and 20 for Low priority). The manager tries heuristically to match the availability of the resources and the number of patients:

- If the number of slots is much higher than the number of patients to invite, the manager moves the invitation expiration deadline to include as many patients as possible without anticipating too much their invitations.
- If the number of slots is not enough, the manager decides if it is necessary to perform overbooking on some priority classes or postpone some invitations to the next month with a time tolerance.

If a reasonable solution could not be obtained despite the heuristic fixes, the manager contacts the pap-test centers asking for additional time availability.

3 Greedy approach

The invitation planning activity, shown in the previous section, relies heavily on trial-and-error, is very error prone, and does not guarantee optimality (or even near-to optimality). Its only chances of success stand in the manager's experience.

We developed two greedy algorithms to support the screening center manager in the definition of the pap-test invitation: Priority-Date and Weighted.

The Priority-Date greedy algorithm schedules the women considering two aspects: the expected invitation date and the priority.

Women in the target population are divided into three different lists depending on their priority. Women in the same priority list are then ordered w.r.t. their expected date.

In each day, available slots are subdivided in three groups according to a percentage associated to each priority level (as described in Section 2). Each group represents the maximum number of slots that can be used for each priority level.

For each priority list, women are extracted from the top of the list and assigned to slots reserved for the corresponding priority. If for some priority the allotted time slots in a month exceed the number of patients of the same priority, the remaining slots are assigned to women of lower priorities.

The Weighted greedy algorithm tries to balance the two aforementioned criteria in order to limit the introduced delays and to give importance to high priority classes. In fact, Priority-Date tends to provide extreme solutions, in which high priority classes are scheduled too eagerly, and low priority patients can be given significant delay.

We give to each patient a weight that depends on her associated delay and priority:

$$W = delay(Patient) \cdot p(Patient)$$

where $delay(Patient)$ is a function that returns the delay of the $Patient$ invitation with respect to the expected examination date and $p(Patient)$ is a coefficient associated to the priority level of $Patient$ (the highest the coefficient, the highest the importance given to the delay). Moreover, as in the Priority-Date algorithm, the user can state that in each day some slots are reserved for patients of a specific priority.

The patients are then ordered according to their weights. Given the ordered list, the algorithm starts the assignment from the first day of the month and associates to a slot reserved for a particular priority level the patient of the corresponding priority with the highest weight. The slots non assigned for this priority level are associated to the women with the highest objective function values independently from their priorities.

3.1 Experiments on greedy algorithms

In order to test the proposed algorithms and highlight their pros and cons, we set up a simulation with very difficult conditions (more women to invite than the available time). The instance spans over 5 months, and involves 2400 women with expected invitation dates randomly generated with uniform distribution. Out of the 2400 women, 1150 were given low priority, 950 normal and 300 high. The pap-test center has a daily time availability of 50 minutes (5 pap-test examinations of 10 minutes or 16 if we consider overbooking with 3 minutes for each examination), 7 days a week.

As shown in Table 1, the Priority-Date algorithm, configured with default parameters (50% of time for high priority, 30% for normal priority and 20% for low priority), gives too much importance to the high priority women introducing significant delays for the low priority women (up to 75 days of delay). The introduction of an objective function in the weighted greedy algorithm represents an evolution of the Priority-Date one, capable of reducing the delays for low priority women (up to 52) as shown in Table 1. It also introduces, for each day, a better allocation of the available slots by balancing priorities and delays in the objective function.

The problem of this greedy algorithm is that it cannot identify an optimal invitation plan as it only discovers local optima. Consider for instance a day in which low priority

Table 1. Max number of delay/anticipation days

Algorithm	Priority	Max Anticipation	Max Delay
Priority-Date	HP	22	1
Weighted	HP	0	16
Priority-Date	NP	5	9
Weighted	NP	0	16
Priority-Date	LP	0	75
Weighted	LP	2	52

patients are subject to overbooking, and a free slot of 10 minutes. We can accommodate either 1 high priority woman with a time delay of one day or 3 low priority women with a time delay of one day each. If we have assigned to the high, normal and low priority levels respectively a weight of 10, 7 and 4 in the objective function. The algorithm orders patients according to their weight: first the high priority woman whose weight is $(10 \cdot 1 = 10)$, then the three low priority patients whose weight is $4 \cdot 1 = 4$ each. Indeed, even if the delay of low priority women rises up to two days they are still ordered after the high priority woman. The weighted algorithm then selects the first patient in the list, assigns the slot of 10 minutes to the high priority woman, thus delaying the three low priority women of one day. This solution costs $10 + 4 \cdot 2 + 4 \cdot 2 + 4 \cdot 2 = 34$.

Looking globally to our list, we observe that the one generated is not the optimal solution as reserving the 10 minutes slot for inviting the 3 low priority women and delaying the high priority invitation of one day has a lower cost $10 \cdot 2 + 4 + 4 + 4 = 32$.

For this reason we used artificial intelligence techniques and Constraint Programming for identifying the optimal invitation plan (the plan that has the lowest sum of all the woman objective function values). This approach is described in details in Section 4.

4 Constraint Programming

The greedy algorithms presented in Section 3 provide reasonable solutions in a very short time. The generated appointment schedules were submitted to the final users, that deemed them acceptable. However, due to the combinatorial nature of the problem, a greedy algorithm in general does not provide the optimal solution, and it never proves optimality.

We decided to experiment with optimization algorithms, in order to find the optimal solution, and to compare the quality of the solutions given by optimal and greedy algorithms. The aim was to evaluate the viability of an Artificial Intelligence module, exploiting a complete algorithm, in the appointment scheduling application.

Constraint Programming (CP) languages are devoted explicitly to the solution of hard combinatorial problems. Initially born as a rib of Logic Programming, CP was then extended also to the object-oriented paradigm. Modern CP languages contain libraries and solvers for different domains. Popular instances are CP(FD), in which the unknowns range on Finite Domains, and CP($\mathcal{R}$), in which variables range on the set of real numbers. The corresponding solvers are based on tree search enriched with prop-

agation algorithms reaching Arc-Consistency (and its generalizations) for the FD domain, and on (Integer) Linear Programming for the domain of the reals.

We first experimented the viability of a CP(FD) model, but it did not provide optimal solutions in reasonable time. We then applied a CP($\mathcal{R}$) solution, exploiting an integer linear programming model, that opens the way to efficient solvers based on linear programming enriched with a branch and bound strategy.

4.1 CP($\mathcal{R}$) Model

At a first sight, one could think to associate a decision variable AD_i, representing the appointment date, to each patient. Unluckily, the number of patients could be large, and many of them share same category and expected date, so the search space can contain an exponential number of symmetric solutions obtained by permuting patients with same features, that gives a well known combinatorial explosion of the search space [4]. Symmetric solutions can be pruned by adding the constraints $AD_i \leq AD_j$ whenever $i \leq j$. However, the number of variables is still very large. Therefore, we decided to classify the patients into groups, each group being identified by a expected date and a category, and associate a variable to each group.

Suppose we have ng groups and nd days. For each group of patients g, and for each possible invitation day id, we define a positive decision variable $I_{g,id} \geq 0$, representing the number of patients from group g invited in day id.

For each decision variable there is a *cost* associated to such assignment. For the group of patients g the cost depends on the category and on the introduced delay with respect to the expected day $ed(g)$. Categories with higher priorities will contribute with a higher cost than low-priority categories. The cost depends on the delay through a nonlinear function. If the invitation date coincides with the expected date, the cost is zero; the same holds if the invitation date is before the expected date, provided that the anticipation is limited: there exists a parameter α defining the maximal number of days a patient can be called in advance. The protocol required delays not to be higher than 40 days; we defined a parameter δ (that defaults to 40). Delays superior to δ or patients called more than α days before their expected date contribute to the total with a very high cost M. A delay between 0 and δ contributes with a cost proportional to the number of days of delay, multiplied to the priority coefficient $p(g)$ of the group g. The objective function is then:

$$min \sum_{g=1}^{ng} \sum_{id=1}^{nd} I_{g,id} cost(g, id - ed(g))$$

were the cost is defined as

$$cost(g,d) = \begin{cases} 0 & if \;\; \alpha \leq d \leq 0 \\ p(g) \cdot d & if \;\; 0 < d \leq \delta \\ M & if \;\; d < \alpha \vee d > \delta \end{cases}$$

The constraints (1) impose that the total capacity of the day is not exceeded. $capacity$ is the total number of minutes available for visits in a given day; $duration$ is the duration of a visit, and it depends on the category and on the day (which enables the user to define detailed policies for overbooking, varying the visit durations).

$$\sum_{g=1}^{ng} I_{g,id}\, duration(g, id) \leq capacity(id) \quad \forall id \in 1..nd \tag{1}$$

Note that an instance could be infeasible if the number of days is not enough to accommodate all patients; in such a case a constraint solver does not provide a solution, but simply returns failure. To provide the manager a reasonable answer also in this case, we avoid infeasibility by introducing an additional day with unlimited capacity and with a high cost M to accommodate all patients.

Each patient should be invited exactly once, stated as constraint (2), where $|g|$ is the number of patients belonging to the group g.

$$\sum_{id=1}^{nd} I_{g,id} = |g| \quad \forall g \in 1..ng \tag{2}$$

Finally, the percentage of time devoted to visiting patients of each category should be respected. Actually, in order to fully exploit the power of the optimizer, the problem should not be too constrained (otherwise, if there are no freedom degrees, the optimal solution boils down to the same solution given by a greedy algorithm). We decided to guide the optimization process toward the specifications of the final user as follows. We ask the user to impose a capacity per day per group of patients, $CapPerc(c, id)$. We check whether the total allotted time for each category is enough for visiting all the patients in that category. If the allotted time is enough, we impose that in each day the number of patients of category c is at most the one specified by the user (3). Otherwise (if the total time is not enough for that category), for each day we impose that the number of patients of category c is at least the one prescribed by the user (4).

$$\begin{array}{l} \forall id \in 1..nd, \forall c \in 1..nc\ s.t. \\ \sum_{id'=1}^{nd} CapPerc(c, id') < \sum_{id''=1}^{nd} duration(c, id'') \implies \\ \sum_{g=1}^{ngc(c)} I_{g,id} duration(g, id) \geq CapPerc(c, id) \end{array} \tag{3}$$

$$\begin{array}{l} \forall id \in 1..nd, \forall c \in 1..nc\ s.t. \\ \sum_{id'=1}^{nd} CapPerc(c, id') > \sum_{id''=1}^{nd} duration(c, id'') \implies \\ \sum_{g=1}^{ngc(c)} I_{g,id} duration(g, id) \leq CapPerc(c, id) \end{array} \tag{4}$$

The model consisting of the objective function, constraints (1), (2), (3), (4) and the integrality constraint for each variable $I_{g,id}$ is solved through branch and bound exploiting a linear relaxation for bound computation. The branch and bound algorithm solves the problem to optimality and proves the solution is optimal.

5 Experiments

We selected a series of experiments to compare the quality and the runtime of the greedy algorithms with respect to the use of the CP($\mathcal{R}$) solver. In the experiments, we used an instance with 204 patients to be scheduled in a period of one month, with random expected day. The patients are divided into three categories: 28 patients HP, 74 NP and 104 LP. The visiting time is 10 minutes without overbooking, while it is reduced to 3

minutes in case of overbooking (only for LP patients). The availability of the screening centre is 50 minutes per day, which is not enough to visit all the patients without overbooking, thus some of the patients have to be be moved to the following month.

In Figure 1 we show the distribution of the difference between expected day and invitation day for each of the categories for the weighted greedy algorithm detailed in Section 3. In abscissa we represent the difference expected day - invitation day, i.e., negative numbers represents anticipation with respect to the optimal invitation date, while positive numbers represent delay. In ordinate, we have the number of patients (for each category) that has such an anticipation/delay. The algorithm gives high priority to high risk patients, which are anticipated, with respect to their ideal date, up to 20 days. Correspondingly, delays are introduced for lower priority patients. This shows that there is room for improvement: intuitively, some of the early patients could be swapped with patients that are delayed.

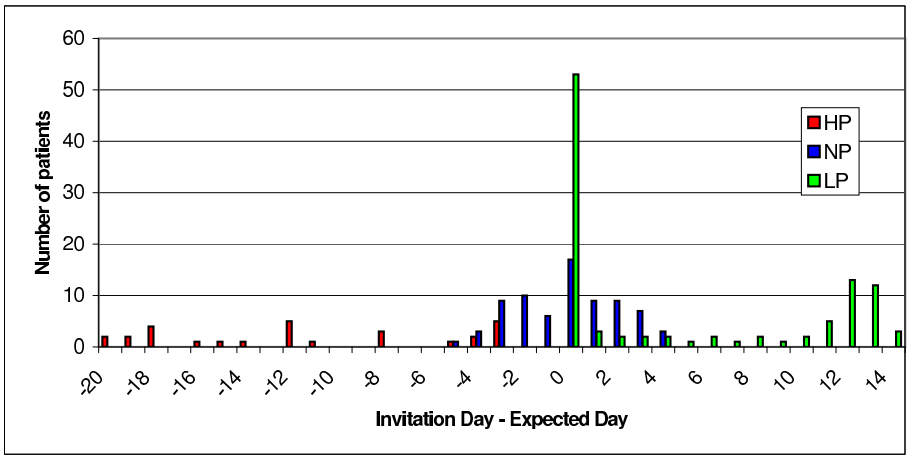

Fig. 1. Distribution of the patients: weighted greedy algorithm

Figure 2 shows the distribution in the optimal solution. Both anticipations and delays are drastically reduced: no patient is anticipated more than 4 days or delayed more than 9 days. The values of objective function in the two situations synthesize the same information visually presented in the graph: the greedy solution has cost 2037, while the optimal cost is almost an order of magnitude better: 325.

The same can be said in the case with overbooking, as shown in Figures 3 and 4. The corresponding costs are 558 for the greedy solution and 153 for the optimal one.

We used ILOG CPLEX 9.0 as solver; it was able to find the optimal solution in a very small time on an Intel Celeron CPU 2.4 GHz, 512MB RAM computer. In order to test the scalability of the algorithm, we experimented with a higher number of patients, up to 20,000. The algorithm scales very well: all the instances were solvable within one minute, which is by far acceptable for an algorithm that is run once every month. The

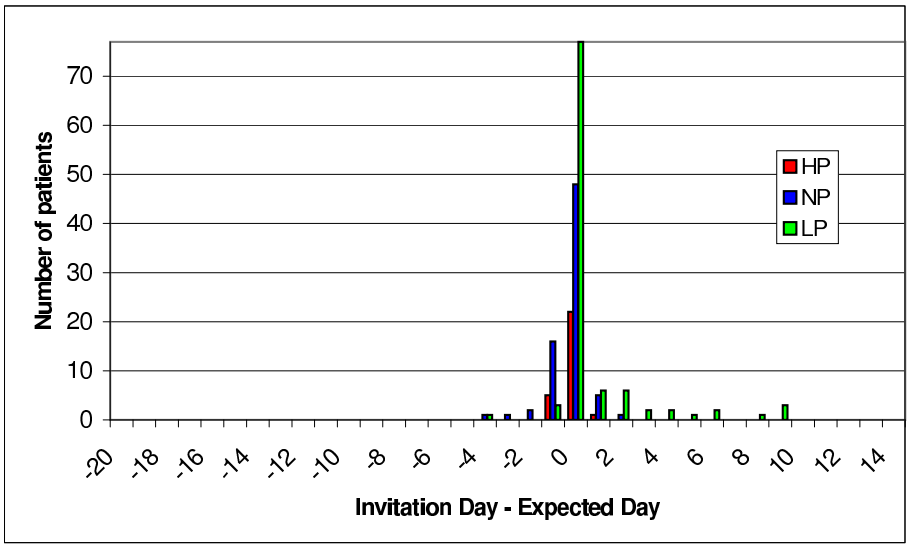

Fig. 2. Distribution of the patients: optimal solution

scalability can be easily explained: the unknowns in our models do not depend directly on the number of patients, which can be large, but the number of groups, that cannot grow beyond the number of possible days multiplied by the number of categories.

6 Conclusions

In the paper we have proposed greedy and exact algorithms for the invitation plan generation for cancer screening.

Invitation plans, generated during experiments performed with different patient and resource configurations, were submitted to the final users, that deemed them acceptable, in any case better than the current hand-generated plans.

Clearly, the choice between a greedy and an optimal algorithm should take into account issues related to scalability, efficiency and solution quality. Small instances (up to hundreds of patients to be scheduled in a month time horizon) can be effectively solved via the exact approach proposed in this paper. When the time horizon raises up to several months we can either face the overall instance with a greedy approach or we can decompose it by dividing the time horizon in monthly slices and solve each sub-instance with the exact algorithm.

Acknowledgments This work has been partially funded by the SPRING (Screening PRotocol INtelligent Government) project, partially financed by Emilia-Romagna Region (Italy) under (PRRIITT 3.1.A), and by the MIUR PRIN project n. 2005-015491. We would like to thank Dr. Natalina Collina of the Sanitary Agency of Bologna for her contribution on the description of the screening process management. We also thank Evelina Lamma and Paola Mello for their useful suggestions.

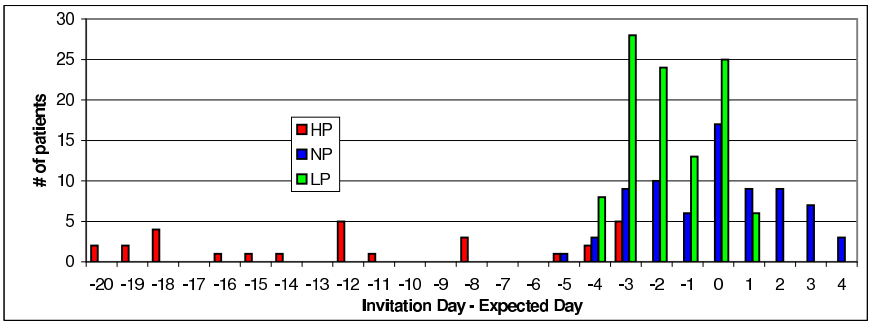

Fig. 3. Distribution of the patients: weighted greedy algorithm with overbooking

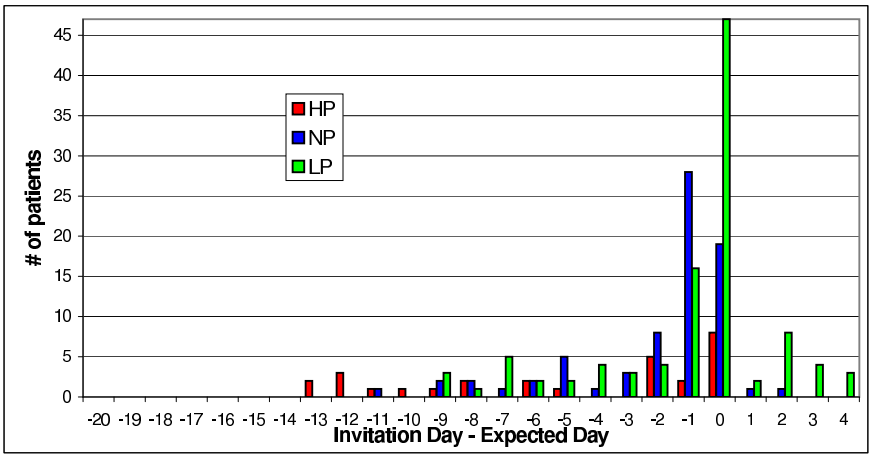

Fig. 4. Distribution of the patients: optimal solution with overbooking

References

1. R. Baker and P. Atherill. Improving appointment scheduling for medical screening. *IMA Journal of Management Mathematics*, 13:225–243, 2002.
2. Cervical cancer screening in the Emilia Romagna region of Italy. Home Page: http://www.regione.emilia-romagna.it/screening/.
3. C. J. Ho and H. S. Lau. Minimising total cost in scheduling outpatient appointments. *Management Sci.*, 38:1750–1764, 1992.
4. J.F. Puget. On the satisfiability of symmetrical constraint satisfaction problems. In *Proceedings of ISMIS93*, pages 350–361, 1993.
5. T.R. Rohleder and K.J. Klassen. Using client-variance information to improve dynamic appointment scheduling performance. *Omega*, 28:293–302(10), 2000.